Head- and Eye-based Features for Continuous Core Affect Prediction

Jonathan (Jonny) O' Dwyer

Athlone Institute of Technology

Submitted as part fulfilment of the requirements for the degree of

Doctor of Philosophy (Ph.D) in Software Engineering.

Supervisors: Dr. Ronan Flynn and Dr. Niall Murray

Submitted to Athlone Institute of Technology, July 2020

Abstract

Feelings, or affect, are a fundamental part of human experience. Arousal and valence make up core affect and have received intense study in affective computing. Speech and facial features have been extensively studied as predictors of core affect. Other indicators of affect include head- and eye-based gestures, yet these are underexplored for affect prediction. In this dissertation, handcrafted feature sets from head and eye modalities are proposed and evaluated in two audiovisual continuous (core) affect prediction experiments on the RECOLA and SEMAINE affective corpora.

In the first experiment, head- and eye-based features were input to deep feed-forward neural network (DNN), along with speech and face features, for unimodal continuous affect prediction. Two proposed head feature sets and one eye feature set outperformed minimum performance benchmarks, estimated human prediction performances, for arousal prediction on both corpora. The more complex head feature set proposed performed second-best overall, after speech, and best from the visual modalities, for arousal prediction. This feature set obtained validation set concordance correlation coefficient (CCC) scores of 0.572 on RECOLA and 0.671 on SEMAINE. For valence, head feature sets performed best from those proposed, and best overall for valence prediction on SEMAINE (CCC = 0.289), however, these sets were unable to match or exceed human performance estimates. From this experiment, it was concluded that head-based features are suitable for unimodal arousal prediction. It was also concluded that arousal prediction performance within -15.82% of speech, relative CCC, can be obtained from head-based features.

In the second experiment, the proposed feature sets were evaluated with speech and face features for multimodal continuous affect prediction using DNNs. The experimentation included a fusion study, cross-modal interaction feature investigation, and the proposal for, and evaluation of, teacher-forced learning with multi-stage regression (TFL-MSR). TFL-MSR is a method for leveraging correlations between affect dimensions to improve affect prediction. An algorithm screening-based sensitivity was also performed to highlight important feature groups for prediction in the different corpora. Model fusion performed better than feature fusion in the experiment. Relative CCC performance increases of 4.91% and 18.23% on RECOLA and 13.18% and 74.17% on SEMAINE above model fusion speech and face were observed for arousal and valence respectively for multimodal systems that used all modalities. One eye and face cross-modal interaction feature was discovered for valence prediction on RECOLA and it was able to improve CCC prediction performance by 2.66%. TFL-MSR improved valence prediction on RECOLA but not on SEMAINE where a small arousal and valence correlation relationship was present. Interesting cross-corpus similarities and differences were found in the sensitivity analysis that indicated some feature groups have similar importances, while other feature groups' importances were inverted across the social situations in the corpora. The final models of this work produced test set CCC results of 0.812 for arousal and 0.463 for valence on RECOLA and 0.616 for arousal and 0.436 for valence on SEMAINE.

The usefulness of the proposed head and eye features has been shown in this research, and they can also facilitate model interpretability efforts as the handcrafted features are themselves interpretable. This work provides researchers with new affective feature sets from video and methods that can improve affect prediction and potentially other social and affective computing efforts.

Declaration

I hereby declare that the work contained in this dissertation has not been submitted by me in pursuance of any other degree. All the work in this thesis is the result of my own investigations, except where otherwise stated. References are given where other sources are acknowledged.

_____

Jonathan (Jonny) O'Dwyer

Acknowledgements

I want to thank the Irish Research Council for their generous scholarships (Grant nos. GOIPG/2016/1572, GOIPG/2018/2030). I also want to thank Dr. Ronan Flynn and Dr. Niall Murray of Athlone Institute of Technology for their support, mentorship and friendship throughout the project. Finally, I wish to thank the Examiners for their helpful comments that greatly improved the final version of this dissertation. Without all of these organisations and people, the development of this document would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Emotion has been studied since ancient Greece and has attracted a lot of scholarly attention. Descartes proposed a number of passions (emotions) of the soul (mind) in his final published work [1], while Darwin postulated that emotion expressions could be deciphered from visual and audible cues in man and animals [2]. In recent times, affective computing has become a thriving research field, with multidisciplinary research efforts being developed from psychology and computer science (among other fields) for automatic recognition and synthesis of affective states. With the increasing development and ubiquity of technology today, there has never been a more important time to extend our understanding of the cues available for affective computing. Affective computing can be a power for good, with applications possible in fields such as cyber (patho)physiology/psychology assessment [3]–[6] and human-computer interaction [7].

Within affective computing, recognition of basic emotions or categories of affect had traditionally been the community's focus. However, there has been increased interest in continuous prediction of dimensional affect in recent times [3], [8], [9]. This is perhaps due to the improved realism of this approach. For example, emotion and affect are inherently subjective and nuanced, transcending clear category boundaries in all but stereotypical or prototypical displays. Moreover, affect varies across time and expression in different people, or even within the same person in a different context [10]. Predicting affect dimensions continuously allows for representations of affect that may escape human verbal description and allow temporal gradients of affect to be obtained [11]. Speech, video, face and physiological measures have been investigated for continuous affect prediction [12]–[15]. However, visual affective cues from head- and eye-based modalities have not been explored to any great extent.

1.1 Problem statement

1.1.1 Problem space

In continuous affect prediction, researchers aim to automatically predict a numerical value for affect dimensions in pseudocontinuous time. In order to make predictions, a model (or models) must be trained using a machine learning (ML) algorithm (or numerous algorithms) on data. An example of the generation process for primary, or raw, data can be seen in Figure 1.1 (a), where two individuals are interacting using computers connected by a network. The subjects are encoding their affective states into an audio-video recording while also decoding their interlocutor's affective states from the audio-video stream as part of the social situation. This recorded primary data is then annotated, as in Figure 1.1 (b), where a human rater provides numerical judgements of their perceptions of affective dimension intensities based on a subject's external display. The annotations are usually provided by more than one annotator [3], [11], [16] in order to gather a more reliable estimate of affect perception for later model training and prediction performance evaluation.

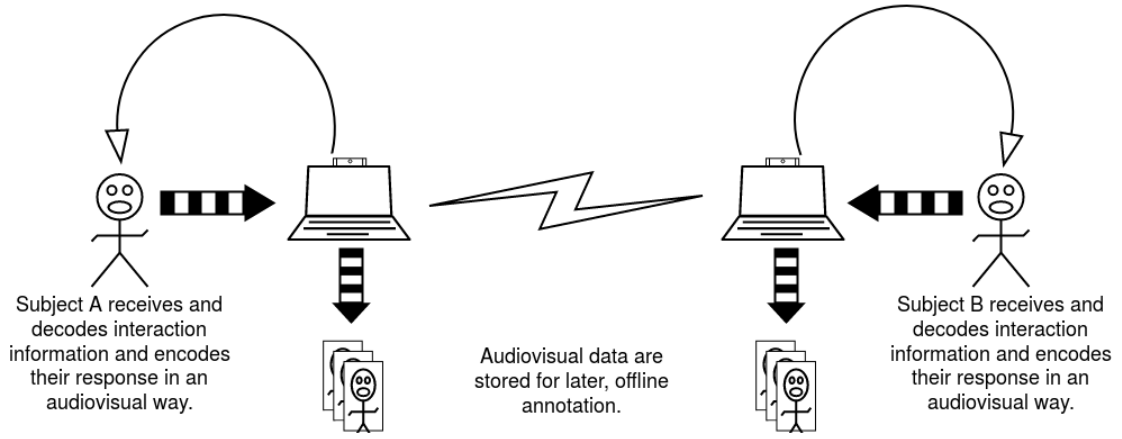
After primary data has been gathered and annotated, features are extracted from the audio-video data. A model is then trained using these features as inputs to the model as depicted in Figure 1.1 (c). Extracting features involves calculating attributes of the raw data. Ideally, these attributes perform well for the prediction task, in this case, continuous affect prediction. The focus of the work presented in this dissertation involves investigating the benefit of head- and eye-based feature sets for audio-video continuous core affect prediction. Core affect is composed of two dimensions, namely, arousal, ranging from activated to deactivated, and valence, ranging from pleasant to unpleasant. It is a component of all affective experience [17] and due to its ever-present nature, successful prediction of core affect is beneficial for affective computing.

1.1.2 Research question

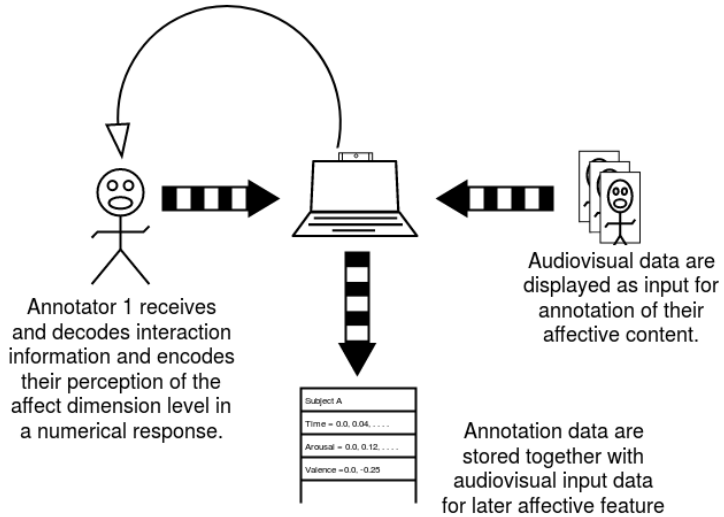
The research question addressed in this Ph.D dissertation is

For audiovisual communication, how much of an improvement in the continuous prediction of core affect can be achieved by processing the combined cues gathered from an individual's speech, head and eyes?

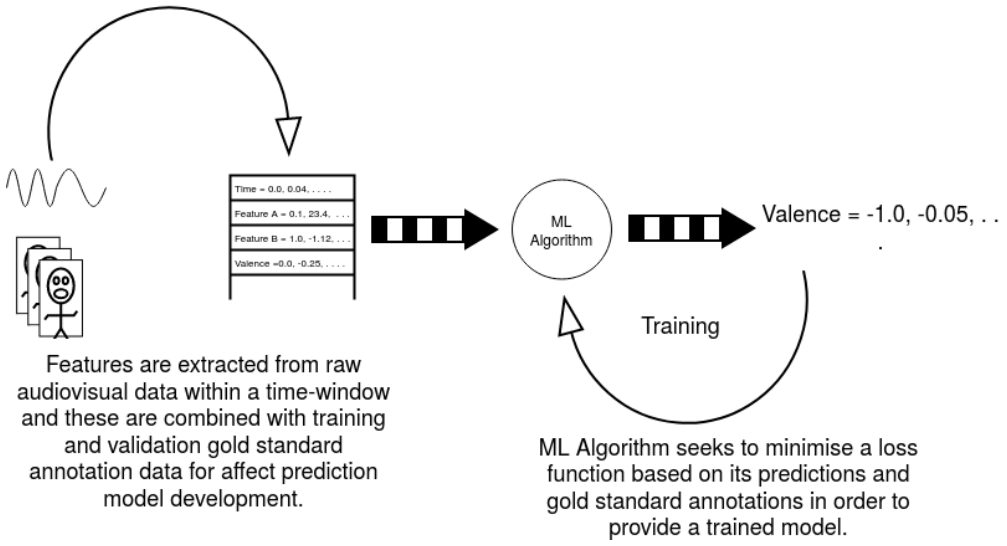
Since core affect [17] is under assessment in this work, references to "affective state" or "affect prediction" for the experiments carried out in this work are used as shorthand for *core affective state* or *core affect prediction* respectively.



(a)



(b)



(c)

Figure 1.1: Continuous affect prediction model development stages: (a) raw audiovisual data collection, (b) affect dimension annotation, (c) affective feature extraction and model training (affect dimension learning) for later prediction.

Multimodal continuous affect prediction aims to increase model performance by leveraging additional information extracted from different modalities. Speech and facial features, for example, have been extensively investigated in affective research, resulting in feature sets and taxonomies [7], [18]–[21] for these modalities. Also, it is known that eye and head gestures can compliment facial displays of emotion [22]–[24], while emotion perception in speech is enhanced in the presence of head-based cues [25], [26]. Furthermore, the eyes' pupils are known to react to affectively salient events [27]–[29], while paying attention to the eyes of individuals can assist in decoding affective signals [30]. Less research attention has been given to cues based on different head- or eye-based measures for continuous affect prediction from audio-video, however, despite evidence of their contribution to affect perception and displays. The head and eye modalities require exploration so that more affective information can be extracted from subject-provided videos and the benefit of considering these modalities as part of affect prediction systems can be assessed. The scope of the research of the visual modalities in this work was limited to their use in the presence of speech with the intention to augment and/or compliment speech for affect prediction. For the experiment chapters of this dissertation, Chapter 4 and Chapter 5, the main research question of this dissertation was further broken into unimodal and multimodal subquestions respectively. This was done to assess the potential of head- and eye-based features on their own and when included as part of multimodal system input. A number of challenges for this research are presented in Section 1.2. These challenges informed experimental designs to sufficiently answer the subquestions and overall research question for this research.

1.2 Challenges

1.2.1 Input features

Engineering a high-quality feature set for modelling algorithms is a difficult task requiring knowledge of the underlying process [31]. As part of the engineering process, a low-level representation of a required feature, a low-level descriptor (LLD), must be sought. In some cases, these LLD representations are calculated from raw or preprocessed raw data and the time resolution for these features often ranges from tens to hundreds of milliseconds. After a set of LLDs has been selected, mid- and high-level features may be obtained across longer time windows, or time chunks [31], to enhance feature representations prior to modelling, resulting in an initial feature set. Some high-level features include mean and standard deviation (SD) functionals of LLDs for a specific time window [21], [32]. It is also possible to generate, or automatically learn, feature representations directly from raw data using

a neural network [33], [34]. Also, the generated features from different modalities may interact with each other (e.g. head movement and speech interaction) resulting in further informative features of affect that should be investigated for multimodal systems [35]. Also, in the process of generating features as input to algorithms, some poor predictors of affect may be added inadvertently. Challenges raised by this process include:

- What features should be gathered from specific modalities and how should they be extracted for the purpose of continuous affect prediction?
- Do cross-modal feature interactions exist and are they useful for prediction?

The Literature Review, Chapter 2, indicates how the first research challenge question can be addressed. Chapter 4 provides a proposal for an evaluation of head- and eye-based feature sets for continuous affect prediction inspired from the reviewed literature. The extraction of the proposed features using different temporal windows is also evaluated in the chapter. The second research challenge question listed above pertains to multimodal evaluation of the proposed features. This issue is addressed in Chapter 5.

1.2.2 Modelling

For modelling continuous affect, considerations can include which modelling technique to use, how to employ features or predictions from different modalities for final predictions and how best to take advantage of correlated affect dimensions. Both traditional machine learning and deep learning modelling techniques have been employed in affective computing. Each modelling technique brings with it challenges such as gathering enough training data, selecting appropriate parameter values and avoiding overfitting in model training. Also, researchers accept that affective dimensions are correlated [35] and this can be used for modelling purposes, however, affect dimensions are often modelled separately [14], [36].

Some challenges of continuous affect prediction modelling therefore include:

- What algorithm or modelling technique is best for the prediction of continuous affect?
- How should features or modality predictions be fused together to make the most effective multimodal model?
- Should correlations between affective dimensions be employed for modelling, and if so, how should this be done?

For the first research challenge listed above, the review of relevant research, Chapter 2, showed what modelling techniques have been used, and research opportunities, for continuous affect prediction modelling. The second and third research challenge questions above are addressed in Chapter 5 of this document, where multimodal experimentation was conducted.

1.2.3 Gold standard annotations

In continuous affect prediction, target values for learning and prediction consist of estimations of consensus human annotator judgements for perceptions of affect based on a subject's external (face, eyes, speech) display. Therefore, a limitation in modelling is that the target to be learned and predicted is not true in a literal sense and the term "ground-truth" does not apply for continuous affect prediction. The annotations provide an accessible, subjective affect perception from a group of raters in pseudocontinuous time and this way of measuring latent affect variables is termed gold standard. Since human annotators provide the gold standard, factors that can have effects on the provided affect judgements include individual annotator (dis)agreement with consensus, affective predisposition/personality [10] and reaction time [9]. Therefore, an important challenge for this work is:

- How should annotations be processed in order to promote gold standard annotation reliability and model performance?

This research challenge is addressed in both Chapters 4 and 5. In each of these chapters, effort to improve the gold standard by way of annotator delay compensation was performed.

1.3 Research objectives

Following from the identified research question and challenges that arise in continuous affect prediction, the research objectives of the work presented in this dissertation are as follows:

- An evaluation of previous work and current approaches related to continuous affect prediction using speech-, head- and eye-based cues and the identification of opportunities for advancing continuous affect prediction.
- The proposal for feature sets from head and eye modalities along with temporal feature window, gold standard time-shifting, and feature selection investigations using the proposed features as unimodal input to deep neural networks.

- The further use of the proposed feature sets to include a fusion study and an investigation of cross-modal interactions of the proposed features and speech and face features as multimodal input to deep neural networks.
- Develop and evaluate a method for leveraging correlations in affect dimensions.
- The selection of the final, developed models from this project for continuous affect prediction that are assessed on the test sets based on the validation set experimentation.

1.4 Research contributions

In meeting the research objectives set out for this work, the following research contributions are provided for the affective computing community:

For the first contribution of this work, head- and eye-based feature sets are proposed for continuous affect prediction. The features were evaluated in a unimodal affect prediction experiment to assess their usefulness and their performances were compared with unimodal speech- and face-based affect prediction systems. Different temporal windows, gold standard time-shifts and feature selection techniques were evaluated for the modalities under consideration as part of the experiment. The results showed that only the head-based features are suitable for unimodal arousal prediction from the feature sets proposed. Feature sets from the head performed second-best for arousal prediction, after speech only, on both corpora used for evaluation, and crucially, these performances were above a required minimum performance baseline. Head-based features also performed best for valence prediction on one of the experimental corpora, but they did not meet the minimum performance baseline estimated for valence for this evaluation.

For the second contribution, fusion of the proposed features with speech and face features was investigated to assess if, and by how much, the proposed features can benefit multimodal affect prediction. Cross-modal feature interactions were investigated and a novel valence learning and prediction method that uses arousal annotations on the training and validation sets and arousal predictions on the test set for valence modelling is proposed. Only one cross-modal interaction feature was found for valence prediction, while a few intra-modal feature interactions were found for arousal prediction. These interaction features improved feature fusion prediction performance for some experimental evaluations. The novel method for valence modelling, teacher-forced learning with multi-stage regression (TFL-MSR), was shown to improve valence prediction performance when there is a correlation between arousal and valence, something that is not always present. The best-performing multimodal models produced in this work included the proposed head- and eye-based features.

These models use a simpler core learning algorithm compared to related research while providing good performance for arousal prediction. This experiment showed that the proposed features can benefit multimodal continuous affect prediction by providing additional quality information sources for prediction.

These contributions were informed by a review of the literature, which revealed evidence for research opportunities in developing head- and eye-based cues for continuous affect prediction. From the review, it was shown that head- and eye-based cues are underexplored for continuous affect prediction and descriptors and methods that should be investigated were identified.

Software to allow the research community to extract feature sets used in Chapters 4 and 5 are made available on this project's GitHub repository[1]. Software generated for experimentation/dissemination in this work is released under the responsible AI licence[2]. This source code licence disallows health and medical issue surveillance and diagnostics without human involvement, for example. Also, all the work in this dissertation was carried out on publicly available audio-video research corpora where the consent of the participant subjects was given.

1.5 Peer-reviewed publications

Publications resulting from this work, for which outputs may have been taken verbatim, include the following:

J. O'Dwyer, N. Murray, and R. Flynn, "Eye-based Continuous Affect Prediction", in 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, Sep. 2019, pp. 137–143, ISBN: 978-1-7281-3888-6. DOI: 10.1109/ACII.2019.8925470.

J. O'Dwyer, "Speech, Head, and Eye-based Cues for Continuous Affect Prediction", in 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), IEEE, Sep. 2019, pp. 16–20, ISBN: 978-1-7281-3891-6. DOI: 10.1109/ACIIW.2019.8925042.

J. O'Dwyer, N. Murray, and R. Flynn, "Head and Eye Features with Teacher-forced Learning for Multimodal Audio-Video Continuous Affect Prediction", planned for submission to International Journal of Human-Computer Studies.

1.6 Overview of this dissertation

The remainder of this dissertation is structured as follows. Chapter 2 provides a review of the literature with a focus on head- and eye-based affect signalling and

[1] https://github.com/sri-ait-ie/Non-intrusive_affective_computing

[2] https://www.licenses.ai/

perception and continuous affect prediction using speech-, head- and eye-based cues. Important background information is provided and research results that informed the selection of the proposed features and motivated this project are discussed. Following the review of related work, Chapter 3 provides information on the affective corpora and experimental approach chosen for this research. The selected audio-video affect corpora are presented, discussed and explored quantitatively. The core affect learning algorithm and a high-level experiment architecture for the experimental results chapters to follow are also detailed in this chapter. In Chapter 4, head- and eye-based feature sets are proposed and evaluated in a unimodal continuous affect prediction experiment. The experimental work culminates in Chapter 5 where multimodal fusion of the proposed feature sets with speech and face features is presented. Cross-modal interaction features and affect dimension correlation exploitation by way of TFL-MSR is also investigated in this chapter. A final summary and concluding remarks for this dissertation are given in Chapter 6.

Chapter 2

Literature Review

This chapter provides background theory and a critique of research relevant to this project. In Section 2.1, background on emotion/affect representation is given along with discussion on identified affective phenomena for prediction in this work. This is followed by a review of evidence promoting and informing the use of head- and eye-based cues for continuous affect prediction in Section 2.2. An introduction to commonly used algorithms and performance measures, and a review of continuous affect prediction work related to this research is provided in Section 2.3. Concluding remarks for this chapter are provided in Section 2.4

2.1 Emotion/affect representation

In order to provide computational systems for affect or emotion analysis, a theoretically sound description of emotion or affect must be obtained. There are multiple schools of thought within psychology as to which of basic emotions [37]–[39], appraisal theories [40], [41], or dimensional affect [17], [42] best describe subjective feeling or emotion representation (i.e. an individual's internal subjective state). Basic emotion theorists posit that emotions can be discretely classed, for example happy or angry, based on observed expressions. Appraisal theories rely on subjective recall (appraisal) of an individual's internal state while dimensional affect describes components of, but not all of, affective experience. Emotion concepts (words) are needed to accompany dimensional measurements along with context and other factors [17]. Figure 2.1 illustrates the different approaches taken in the field, ranging from basic emotion to psychological constructionist theories. Each of these approaches have advantages and disadvantages, which are briefly explored in the following sections.

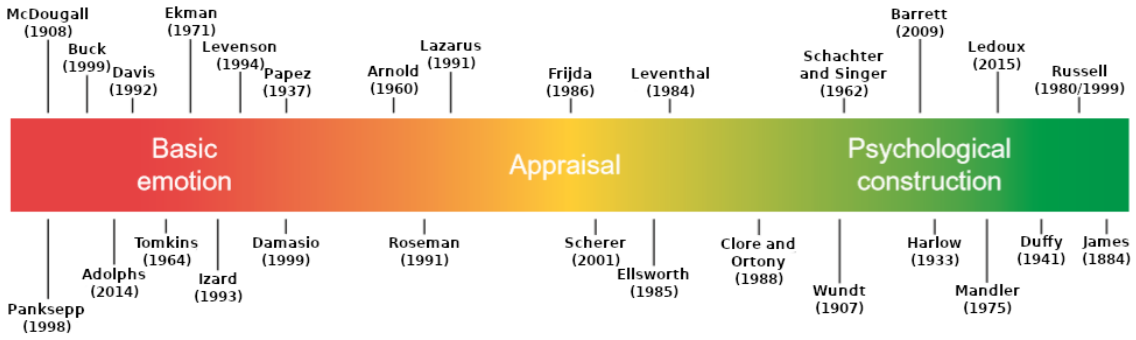


Figure 2.1: Emotion/affect theories ranging from basic emotions through appraisal theories to psychological construction/constructionist from [43]. Image adapted to suit this dissertation with permission. ©Elsevier 2016.

2.1.1 Basic emotion theories

Darwin's theory of emotion [2] suggests that emotions are of a natural kind; they are hardwired, innate, and have universal expression across cultures and in some cases, species. Some basic emotion theorists inspired by this theory include Tomkins [44] and Ekman [38], who base their theories on facial expressions of emotion categories. Other Darwinian examples include Panskepp [37] and Adolphs [45], who argue that emotions represent functional states of the brain with homologies between human and non-human animals. These theorists do not agree on the number of basic emotion categories or functions, but they do all adopt an evolutionary view. Within this framework emotions are viewed as natural, caused, and biologically primitive.

Perhaps the most widely accepted of these emotion theories within affective computing is Ekman's six basic emotions [38]: happiness, sadness, anger, fear, surprise and disgust. An advantage of Ekman's approach is the simplicity of the model, which is attractive for computation as there are a few clearly distinguishable categories to be recognised. Also, intuitively, it might appear that some facial expressions of emotion contain more general information across cultures, for example, prototypical displays of happiness. A drawback of this approach, however, is the potential of basing emotion categories *solely* on facial expressions, whether implicitly or explicitly. Ekman himself acknowledges that emotions can exist without them being facially expressed (e.g. shame and guilt), and that deceptive displays of emotion from the face are possible [46]. More generally, describing emotion solely using basic emotion theory (i.e. discrete, general/natural categories) means reducing a complex event, into a simpler, irreducible one which may not account for all the variation in emotion episodes. This has been contested by appraisal [47] and dimensional theorists [17] within psychology.

2.1.2 Appraisal theories

Appraisal theories rely on subjective evaluations provided by individuals that relate to their own perceived internal state based on events and situations [41]. Compared to basic emotion theory, which suggests that emotion is biologically primitive and innate, appraisal theories require cognitive evaluations of subjective experience. These theories therefore provide for heterogeneity in emotion episodes. Within these theories, there exist both causal (i.e. emotions are caused or elicited) and constructionist (i.e. individuals psychologically construct their emotion or affect) appraisal theories [43]. For a causal appraisal theory example, Scherer's component process theory [41], [48], shown in Figure 2.1, proposes that stimulus evaluation checks linked to valence, activation and power, correspond to relevance, implications, coping potential, and normative significance appraisal objectives. These sequential appraisals determine the emotion that an individual experiences at a given moment. Some appraisal variables within this framework include, novelty, intrinsic pleasantness, and goals/needs for the *relevance* appraisal objective and control, power, and adjustment for the *coping potential* appraisal objective. This model is comprehensive in its approach as it takes the physiology, action tendencies, motor expression, and subjective feelings of the individual into account for each appraisal process as shown in Figure 2.2. On the other hand, the Ortony, Clore and Collins (OCC) appraisal model [40], [49] does not view emotion to be caused by subjective appraisal, instead to be psychologically constructed based on individuals and situations. In this model emphasis is placed on situations and the appraisals provide structure rather than cause. OCC theory appraisals are provided based on valenced (evaluative) reactions to *event outcomes*, *acts of agents* and *aspects of objects* at the highest level, as shown in Figure 2.3. This results in pleased or displeased, approving or disapproving, and liking or disliking affective reactions, respectively, to these differing aspects of appraisal [49].

Self-reported appraisals have been described as possibly the best indicator of prototypical emotion episodes by dimensional affect theorists [17]. Also, subjects implicitly take individual differences and context into account in their appraisals of their own state which is advantageous for this approach. However, appraisal theories do present some practical disadvantages for certain computational applications. The reliance on self-reporting may not be desirable if detailed temporal resolution of affect measurements are required, or if it is either impossible or unreasonable to query a subject for their appraisals. Some pathologies, such as dementia or dysarthria, and computer usage such as gaming or audio-video calling, are some examples. The work of appraisal theorists is important for affective computing, however. These theories provide agreement on the utility of evaluative (valence) measurements [40],

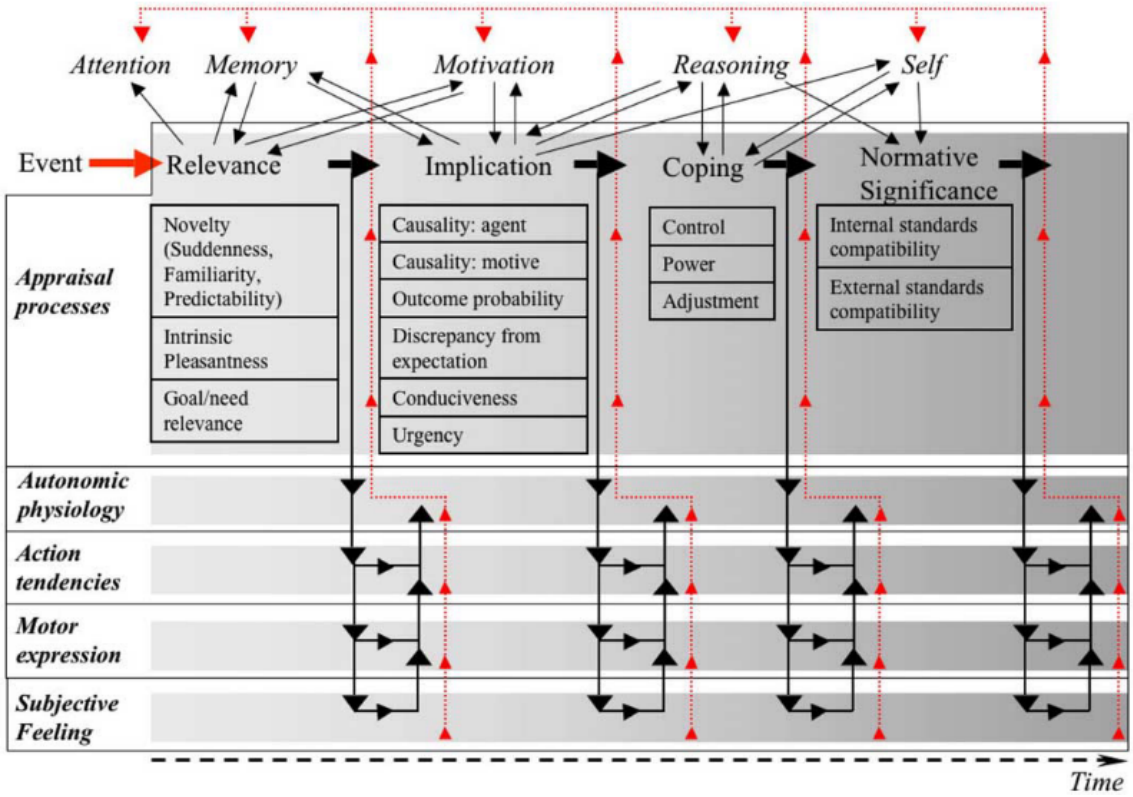


Figure 2.2: Scherer's component process theory from [48]. Copied with permission. ©Elsevier 2005.

[41], [48], [49] for affective computing efforts. Moreover, these theories provide validity for some dimensional theories of affect in addition to highlighting some variables and processes that are beneficial when researchers can take advantage of subjective appraisals.

2.1.3 Dimensional theories of affect and emotion

Modern dimensional theories of affect and emotion follow from Wundt [50], where he reasoned that fundamental feelings, which can develop into emotions, are described as having some mixture of the components: *pleasurable-unpleasurable*, *exciting-depressing*, and *straining-relaxing*. Other components that he claimed as characteristic of emotion include *intensity* and a temporal component called *occurrence*. In recent times, there has been many affect dimensions employed for affect and emotion description, e.g.: *approach, arousal, attention, certainty, commitment, control, dominance, effort, fairness, identity, obstruction, safety, upswing, liking, novelty, intensity and valence* [11], [42], [51]–[53]. Within affective computing, arousal and valence are the most commonly incorporated dimensions in continuous affect prediction corpora despite an argument that emotion is not two-dimensional [51]. The authors of [51] do concede, however, that the additional dimensions that they pro-

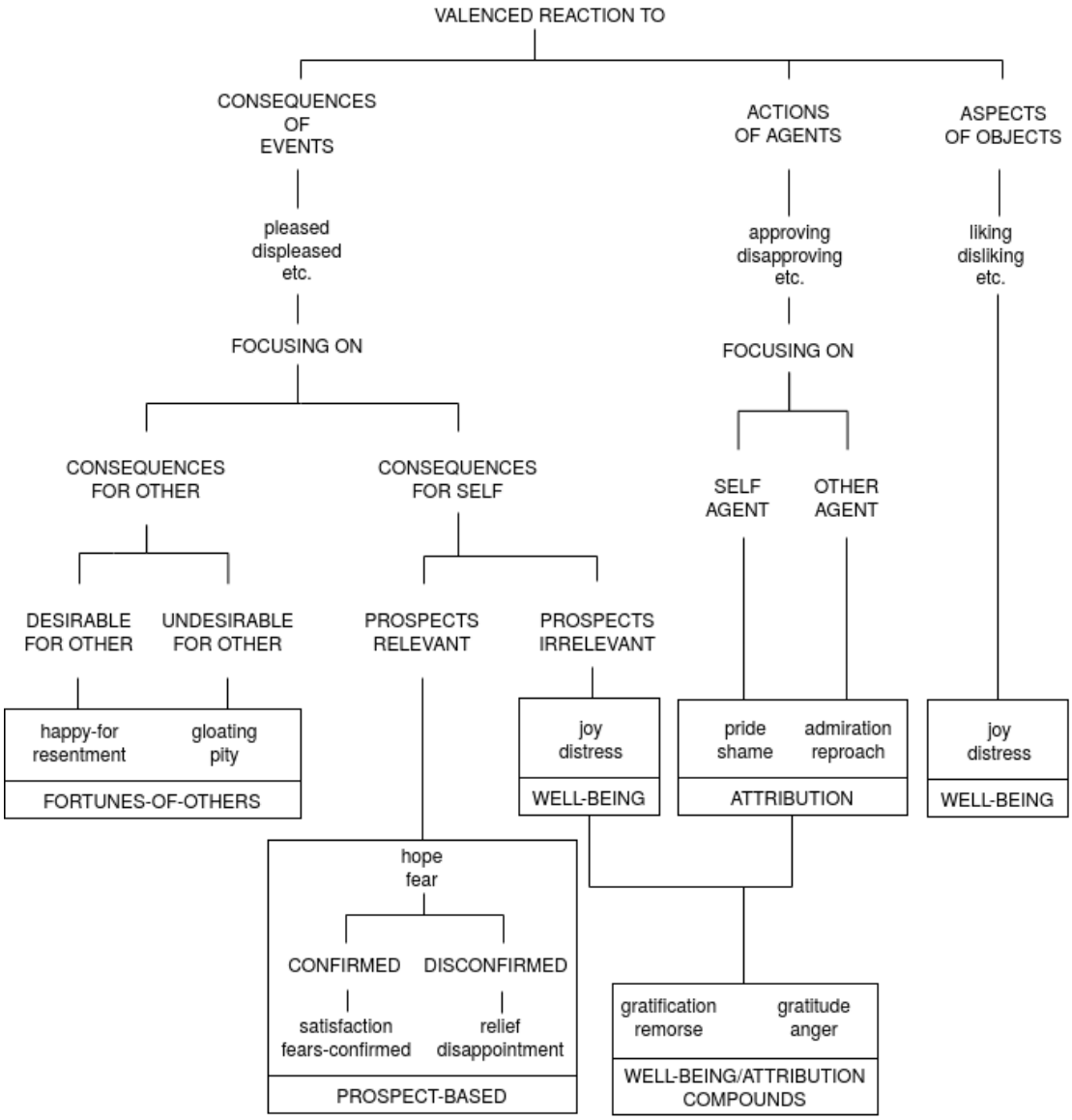


Figure 2.3: OCC model of emotion structure adapted from [40].

pose for emotion description (dominance, novelty) with arousal and valence may not account for all emotion. This means that a low-dimensional representation of emotion, if possible, is still an open problem.

Scholsberg first proposed a two-dimensional structure of facial emotion expression [54] and this model was later refined and shown to represent a cognitive structure of affect by Russell [42]. Russell's circumplex model (Figure 2.4) is composed of two graphically orthogonally placed bipolar dimensions, namely, arousal, ranging from positive (activated/aroused) to negative (deactivated/sleepiness), and valence ranging from positive (pleasant) to negative (unpleasant/misery). This two-dimensional structure of affect was later termed *core affect* [17], which the authors state is always present but need not be directed at anything. Core affect describes important *features* of a prototypical emotion episode. The authors distinguish a prototypical emotion episode from core affect as "a complex process containing an antecedent, appraisal, physiological, affective, and cognitive changes, along with behavioural responses and self-categorisation" [17]. Therefore, core affect is a required but not sufficient phenomenon for measurement in the explanation of complex emotion structure. A core affect measurement can be taken as the distance from the centre of the circle, comprised of an arousal and valence measurement, to a point within or along the circumplex shown in Figure 2.4. A disadvantage to this approach is that it describes only features or properties of an emotion episode as opposed to appraisal theories, which can theoretically provide better estimates of such complex events. Also, this measurement model on its own may be unable to distinguish between combinations of high positive arousal, low negative valence states, for example, emotion instances of fear vs anger. However, core affect does have the advantage of being ever-present, resulting in potentially increased temporal resolution for this type of affect if it can be estimated accurately. Additionally, core affect has the benefit of fewer requirements for measurement as it comprises only one important aspect of emotion, not the entire multidimensional emotion space. These benefits are particularly important for continuous affect prediction where temporal gradients of affect are sought.

2.1.4 Discussion

Both appraisal and dimensional theorists agree that emotion is complex and multifaceted [17], [40], [41], [51]. Therefore, while it is perhaps true that certain prototypical emotion instances may be considered innate or basic, these instances may not always be observable in everyday emotion displays. These appraisal and dimensional theorists additionally agree that both valence (evaluative) and arousal (physiological) features of experience should be considered for affect or emotion assessment.

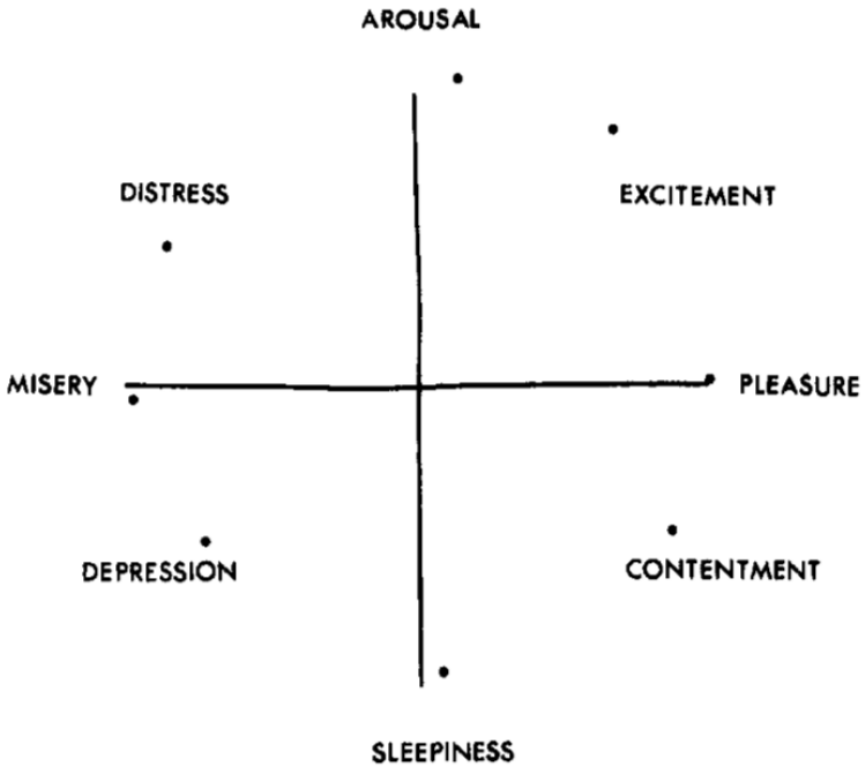


Figure 2.4: Two-dimensional circumplex model of affect from [42] with various affect concept words plotted.

In some cases there is an additional focus on power/dominance (level of control or helplessness) and novelty (also encompassing unexpectedness/unpredictability) dimensions [41], [51]. It is therefore clear, from this review of psychological literature, that the accurate measurement and prediction of valence and to a lesser extent, arousal, is of utmost importance to the affective computing community. The OCC appraisal theory places less focus on arousal [49] for affective experience and emotion, a sentiment echoed in [10]. However, omitting arousal would render affective computing researchers not measuring core affect, which, this author believes, is a fundamental consideration in affective computing research, based on the reviewed literature [17], [40], [41], [50].

In terms of impact, the effective measurement and prediction of arousal and valence can be beneficial in several ways. Arousal, the state of interoception or internal physiological evaluation, interacts with autonomic nervous system activation and wakefulness states of the brain in addition to affect [55]. Its effective prediction can find application in everyday scenarios where high negative arousal (e.g. sleepiness) is unsafe or unwanted, for example, driver drowsiness detection or education delivery assessment. Conditions of hyperarousal are associated with attention-deficit hyperactivity disorder, bipolar disorder and anxiety disorders [56], which allows for

pre-clinical/clinical impact of its successful prediction. Valence, which is also known as exteroception or external evaluation, lets one know how they are doing: do they feel positive/negative, or did they find stimuli pleasant/unpleasant, and to what degree? Excessive negative experience of this affect is associated with both depression and anxiety disorders while sensitivity to this affect is associated with neuroticism [10], [56]. It is clear that the affective computing community acknowledges the importance of arousal and valence, with equal emphasis placed on each dimension in contrast to OCC appraisal theory, for example. Every accepted continuous affect prediction corpus available contains estimations of these measures for affect prediction in subjects [7], [11], [16], [53]. Therefore, effective prediction of these measures provides applicability in a number of contexts as data are available for social situations ranging from human-virtual agent interaction [11] to advertisement watching/discussion [53].

This section has introduced some models for emotion and affect representation that have informed the selection of affect measures for investigation in this research. Inspirational works from psychology have been briefly reviewed and areas of agreement (e.g. valence) identified, while potential areas of impact for affective computing have also been discussed. For the interested reader, further details on appraisal emotion theories related to affective computing can be found in [57], while interesting discussions on basic emotion theories vs constructionist theories are provided in [43] and [58]. The next section deals with the review of evidence promoting head- and eye-based cues as visual descriptors of affect.

2.2 Head and eye-based affect

Despite a large body of evidence linking head- and eye-based cues to emotional and motivational state conveyance [22]–[28], [32], [59]–[65], the use of these cues is underdeveloped for continuous affect prediction. Head cues have been shown to contribute to emotion signalling during speech [24]–[26], [66] and visual displays of emotion [24], [59], [60]. There is also an inherent relationship between head pose and eye gaze [59], [67]. Further, direct (or opposing this, averted) eye gaze is a social signal [22], [23], [68], while the pupils are responsive during emotional arousal [27], [28], monetary incentive, penalty, or after verbalisation [29]. Finally, while eye blink can be considered a confounder for gaze and pupil information, blink rate contributes information on cognitive processes [69] and complete eye closure alters the neural response of perceivers of faces [68]. It is clear that head- and eye-based cues are important for subjects under study. They are also important in the decoding of affect by raters who observe these visual features. The remainder of this section

briefly reviews related research and findings regarding salient descriptors of affect from these modalities.

2.2.1 Head affect

The head can offer important social interaction cues. For example, acknowledgement cues can take the form of nods, while conveying affective states such as embarrassment can be achieved by turning one's head slightly away and downward from an interlocutor according to Cohn et al. [59]. It is usually measured in terms of 3 dimensions, namely, rotation/orientation: yaw; pitch; roll, while a further 3 dimensions are also possible, translation/location: x; y; z, as are higher-level features such as head nods. Kaliouby and Robinson [70] used head nod, tilt and shake HMM features learned from head pitch, roll and yaw sequences respectively. They combined these features with facial gestures to learn mental state classifiers and found that head nod was linked with the agreement state and head shake was linked with the disagreement state. The authors further used the head and facial features combined to successfully recognise complex mental states automatically. This is important for affective computing because the brain may not respect the line between affect and cognition [71]. Gunes and Pantic [72] used similar head features to [70] and achieved results comparable to speech for unimodal continuous affect prediction using SVR. In a cross-cultural perception study by Shibata et al. [60], high-arousal was found to be associated with an upright head and trunk while low-arousal was associated with the trunk moved back and the head moved forward or back. Negative-valence was related to a head and a trunk lying forward across cultures in this study. The authors further showed that trunk and head angles were significant features in regression analysis (arm and leg configurations were also included as regression features) for each of arousal, valence and dominance affect classes. Adams et al. [24] found the head and face to be complimentary during non-verbal subjective emotion recognition in videos. They also showed that features outside of head nods and shakes may be beneficial for emotion recognition systems. Recognition rates did not significantly differ (analysis-of-variance) in videos containing nod/shake cues compared to ones containing general head motion or static head poses.

In terms of cross-modal complimentarity, head pose/gesture is important for use with speech cues. In a study of 17 intimate couples, head pitch angular displacements were found to be larger during conflict when compared to non-conflict discourse by Hammal and Cohn [62]. Busso et al. [25] showed head cues and speech prosody to be strongly linked by objective measures (0.69 average r across emotions). They additionally carried out a subjective experiment that showed emotion perception changing in the presence of different head motion patterns. Livingstone and Palmer

[26] concluded that vocalists' head movements encode emotional information during speech and song, and that observers could identify emotion based on head movement alone. Positive vocalisation was associated with an upright held head while negative vocalisation was associated with a downward held head in their study. Low-level head-based features have been shown to be effective for unimodal emotion prediction in the presence of speech by Ding et al. [65]. Forty-five features based on discrete Fourier transform of angular head movements, gathered using a motion capture system, along with a static measure of head pitch, were used as input to machine learning classifiers in [65]. The best results achieved among the classifiers included 94% for neutral, 79% for sad, 57% for happiness and 72% for angry which indicated that low-level (i.e. not directly discriminating nods, shakes) head-based measures can be useful for emotion classification. Previous research clearly shows that head poses and gestures should be considered for continuous affect prediction, both in the presence and absence of speech.

2.2.2 Eye affect

Eye gaze, the line of sight between an individual and an object of fixation, can be a powerful social and affective cue. While very complex overall, some simple examples of gaze-related movements are: fixational movements (fixation, drift), saccades (faster jerky movements), and smoother-type pursuit movements [67]. The eyes act as both decoders and encoders, with gaze allowing individuals to orient themselves/their attention toward salient stimuli or decode threat cues. Directly gazing toward someone else's eyes while in a state of anger (mutual gaze) allows one to convey dismay directly at the intended subject. This phenomenon has been proposed as the shared signal hypothesis by Adams and Kleck [22], [23], where they suggest that a person's eye gaze shares information with the intended emotion display. Within the shared signal hypothesis, direct gaze toward a person is said to be related to approach-oriented emotions such as joy or anger, and averted gaze related to avoidance-oriented emotions such as sadness or fear. The eyes have also been shown to be important features when processing faces during eye saccade simulation [73] and in the perception of facial emotion displays [22], [23], [74]. Direct eye gaze has been shown to contribute to attentional blink [63], which is to say, increased attentional processing on direct gaze stimuli that can cause reduced or diminished attentional processing of stimuli at later points in time. Also, Duncan and Feldman Barrett offer the hypothesis that the amygdala enhances visual awareness, ensuring environmentally salient information reaches conscious attention [71]. This is in contrast to traditional thinking of the amygdala as a threat/fear detector. More recently, interpretation of a study with a patient who suffered amygdala damage,

[30], indicated that it may not be that the amygdala is a fear detector. Instead, the amygdala may facilitate spontaneous attention to salient cues associated with fear, in the case of [30], widened eyes.

In affective computing research on eye gaze, a direct/averted gaze feature was used to improve facial expression emotion recognition in images by Zhao et al. [75]. Their results showed that angry, sad, fear and disgust recognition could be improved by considering direct/averted gaze with facial features. Ringeval et al. [76] attempted to automatically predict both subject-provided annotations and depicted audio-visual displays of affect, based on the annotator's eye gaze data as input to SVM. The authors additionally provided correct and random emotion feedback (using an on-screen emoticon) to the annotators as part of their experiments and used gaze count (gaze fixated on emoticon), gaze interval (time spent away from emoticon) and horizontal and vertical eye gaze movement as their LLDs. The authors calculated statistics based on the LLDs as part of their feature set and recognised passive/active arousal and negative/positive valence ratings provided by the subjects 82.2% and 69.9% in cross-validation F-scores respectively. Of further interest in this work is that the authors found an improvement in classifying the actual, depicted and gold standard valence values based on the gaze input data, a score of 74.8%. This work showed that eye gaze data of perceivers may be useful in generating objective valence annotations. Wang et al. [77] used a Tobii EyeX controller to gather on-screen x, y eye gaze data that was used together with computer mouse features for stress recognition in computer users. The feature vector contained gaze-mouse coordination features, including average speed of gaze, speed of mouse and others. Correctly classified rates of 94.4% and 82.9% were achieved in 5-fold cross-validation and LOSO-CV respectively, showing the usefulness of eye gaze and mouse cues in stress detection.

The human pupil allows light to enter the eye retina. Its size is controlled by the muscles of the iris, which contains nerves and receptors for the autonomic nervous system, known to generate response output under numerous emotional states [78]. While the pupil size is known to vary under environmental, pathological and pharmacological conditions [79], there is a body of evidence suggesting its efficacy for outward neuropsychologic and affective signalling in healthy individuals as well. It was demonstrated that the pupils provide parasympathetic and sympathetic nervous system signals by Franco et al. [61]. The authors applied physical stimuli (light flash, cold) to subjects while measuring pupil frequency responses using singular spectrum analysis and wavelet analysis (Daubechies family of order 10). The results obtained showed characteristic frequencies for parasympathetic and sympathetic nervous system activity from the stimuli, with wavelet analysis achieving comparable results to singular spectrum analysis. Neuropsychological evidence has been provided for

pupilliary responses to reward expectation (forecasted positive valence event) [64]. The pupils have also been known to reflect cognitive load for some time [80], [81]. An EyeLink 1000 eye tracking device was used by Aracena et al. [82] to gather pupil size and gaze measurements from individuals observing image stimuli and a decision tree neural network was used emotion recognition based on those inputs. The model classified the positive/negative/neutral responses of individuals correctly at a 53.6% rate on a subject-independent basis, based only on pupil size and x, y gaze temporal sequence features.

Partial eye opening or closure events are involved in certain eye gazes, saccades [83] and facial expressions of emotion [18]. Additionally, it was suggested that eye gaze and blink share common signalling pathways by Engelke et al. [84], which makes incorporation of eye closure and blink important for a complete investigation of potential eye-based features for affect prediction. Perhaps the most comprehensive work in using eye-based cues for affective computing is that of Soleymani et al. [32]. Their features included statistics and spectral power calculations from: pupil diameter, gaze distance, eye blink, x and y gaze coordinates, and eye scanning and fixation gathered using a Tobii X120 eye tracker. Their eye-based features performed best when compared to electroencephalogram and peripheral physiology measures as input to a SVM for arousal and valence low/neutral/high classification. For their multimodal experiment, bimodal fusion of electroencephalogram and eye-based features performed the best overall (arousal = 67.7%, valence = 76.1%) in LOSO-CV. The results in [32] indicate that eye-based cues are worth investigating for affect prediction systems.

2.2.3 Head and eye affect

The head and eyes share a close relationship. For example, shortening one's gaze assumes head location displacement. Another example of this relationship include certain emotion signalling of the approach-oriented emotion, joy, being accompanied by a person's head being held more upright [26], with gaze directed toward the joyous stimuli [22], [23]. Alternatively, for avoidance-oriented emotions such as sadness or fear, a person's head may be rotated lower [26] along with an averted gaze away from the threat or unpleasant stimuli [22], [23].

In multimodal affective computing research incorporating head and eyes, head pitch was shown to move downward, along with changes in yaw and roll, with gaze moving slightly upward during embarrassment smiles by Cohn et al. [59]. In automatic frustration recognition by Kapoor et al. incorporating eye blink, posture, head and computer mouse pressure measures, it was found that face and head fidgets and head velocity were the most discriminative features [85]. However, the authors note

that fidget measurement may not have been as reliable due to outlier values observed. Ramirez et al. [86] used Omron OKAO vision software to obtain horizontal and vertical eye gaze, head tilt and smile intensity from video to be used as visual input feature vectors. These visual features of dimensionality 4 performed comparably to the AVEC 2011 baseline visual feature set [87] of dimensionality 5,900, each using a SVM for high/low arousal and valence classification. They provided additional validity for their feature set by outperforming the challenge baseline on the test set using a temporal classification model. Wu et al. [88] used a single high-level head pose and eye gaze cue which they showed improved arousal and valence prediction when combined with facial features, compared to unimodal facial-based affect prediction for continuous affect prediction.

Respondent reactions to negotiation offers were predicted in a dyadic scenario using a multimodal system (which included speech, eye gaze, head pose and smile features) at a rate of 70.8% on average for cross-validation by Park et al. [89]. The authors further noted that symmetric smile, posture, head pose and eye gaze were predictive of negotiation acceptance, while asymmetric head pose and eye gaze were predictive of rejected proposals. Head- and eye-based cues have also been incorporated into the innovative computational behaviour prediction framework MultiSense by Stratou and Morency [4]. This framework includes open-source and commercial software and achieved $r = 0.882$ for distress prediction in LOSO-CV across 100 subjects. Eye gaze x, y and z, gaze distance, along with 6 degrees of freedom head pose were incorporated into their multimodal framework. Eye opening, gaze, averted gaze, blink, and head yaw, pitch, and roll were used with speech for depression classification by Alghowinem et al. [5]. It is also noted that the AVEC 2017 [90] depression challenge included head pose and eye gaze features in baseline sets. It is clear that the affective computing community is embracing head- and eye-based cues, particularly within the psychopathology domain. These are important steps toward fully understanding and utilising all available data from subject videos and, moreover, improving affective computing performance and outcomes for society.

2.2.4 Discussion

It is clear from this review that head- and eye-based cues contain useful social and affective signals. Healthy humans are adept at processing these visual signals for disambiguating affective displays [22], [23], [30], [71], [74]. Researchers have successfully incorporated head-based cues in terms of high-level nods, shakes and fidget cues [72], [85], [86] and low-level head measurements in 3 or 6 degrees of freedom [4], [25], [26] for affective computing tasks. Low-level head measurements have the potential to capture cues outside of head nods and shakes, and they have

been shown to be useful in emotion recognition [24], [65]. Eye-based cues for affective computing can be comprised of the pupils, gaze, saccades and numerous forms of eye opening and closure [5], [32]. A further knowledge-based cue such as direct (and consequently, averted) gaze can be beneficial for automatic systems [5], [75], while features within gaze such as fixation, scanning and eye closure (lack of gaze) can also be useful extracted features [32].

A drawback of using some head- and eye-based cues together is that they may be correlated, perhaps due to their inherent relationship. However, they may offer complimentary signalling for some affective states [59], [88], [89]. Looking at cross-modal correlations may be interesting to understand correlated features but it is not yet certain how these correlations (or lack thereof) interact in some commonly used deep learning systems. Further, inspired by Gunes et al. [35] who identified cross-modal feature interactions as important for investigation, interactions of head- and eye-based features should be explicitly explored. These interactions could include, as first steps, additive, subtractive, product or quotient interactions. The current literature implicitly looks at feature interactions, for example, by early feature fusion before providing input to an algorithm. Explicit feature interaction can provide new features and insights into affect signalling and perception.

This author acknowledges that pupil cues run the risk of measuring nothing particularly well if they measure everything (e.g. light accommodation, pharmacological, pathological, and affective/cognitive). However, their inclusion is important for a complete study of eye-based cues for affect prediction and there is evidence of their efficacy in providing affective signals [61], [64], [80]–[82]. Frequency-based measures should be incorporated from the pupils [61], perhaps in the form of short-time Fourier transform or wavelet time-frequency representations. Overall it is believed that head- and eye-based cues can offer complimentary features for multimodal continuous affect prediction in audio-video. Further, it is believed that head-based cues can offer competitive performance compared with speech for unimodal arousal prediction in audio-video. Clear advantages for the head-based modality for this task are a lack of (audible) noise and reverberation, disadvantages include noises in the visual domain and illumination. Eye-based cues are important to understand fully due to their dual role in the affective computing process, both as signalling components in subjects and important decoding (perhaps even objectively so [76]) components for annotators of audio-video. These cues can be estimated non-intrusively from video using open-source innovations in computer vision such as OpenFace [91], [92], which provides ease of use for researchers.

Also, it should be mentioned that there are ethical implications for the research of head- and eye-based cues for continuous affect prediction, in spite of the benefits that they can provide. As much as facial pose/gesture, the proposed modalities

can offer alarming negative consequences of unwanted affect mining or unlawful affect profiling. Examples of these affective computing mal-uses respectively include unauthorised affect prediction or profiling someone as depressed without due clinical assessment. In the former case what is particularly sinister behind this mal-practice using eye and head cues is that these cues are not commissioned in the same way as speech. Trying to remain affectively silent in this context means closing one's eyes and not moving one's head or blocking the camera. Clearly, this is not always feasible with the number of computers and applications that make use of cameras in our daily lives today. In line with these concerns, and standard ethical research practices that are also adhered to in the work presented in this dissertation, software generated for experimentation/dissemination in this work is released under the responsible AI licence[1]. This source code licence disallows health and medical issue surveillance and diagnostics without human intervention, for example. Also, all of the work in this dissertation was carried out on publicly available audio-video research corpora where subjects consented to filming and analysis and interact in an audiovisual way. While claims can be made about the visual modalities in this work, they are in the presence of speech and intended to augment and/or compliment it. Therefore, performance cannot necessarily be extrapolated to *completely* visual affect prediction such as the aforementioned mal-use where the only "off switch" is to block the head and eyes or camera.

2.3 Continuous affect prediction

Continuous affect prediction can allow for representations of affect which may escape human verbal description and allows temporal gradients of affect to be obtained [11]. This form of prediction can enable both short-term transient (state) and longer term (trait) detection of affective states which can be of benefit in numerous domains. This section provides details on commonly used algorithms and performance measures before critical review and discussion of continuous affect prediction literature. Specifically, SVR and long short-term memory recurrent neural network (LSTM-RNN) learning algorithms and the concordance correlation coefficient (CCC) performance measure are presented and discussed. This is followed by review and critique of state-of-the-art continuous affect prediction research. The aim of this section is to present accepted methodologies and open research opportunities applicable to this research. The focus of this section is continuous arousal and valence prediction in audio-video using speech-, head- and eye-based cues.

[1] https://www.licenses.ai/

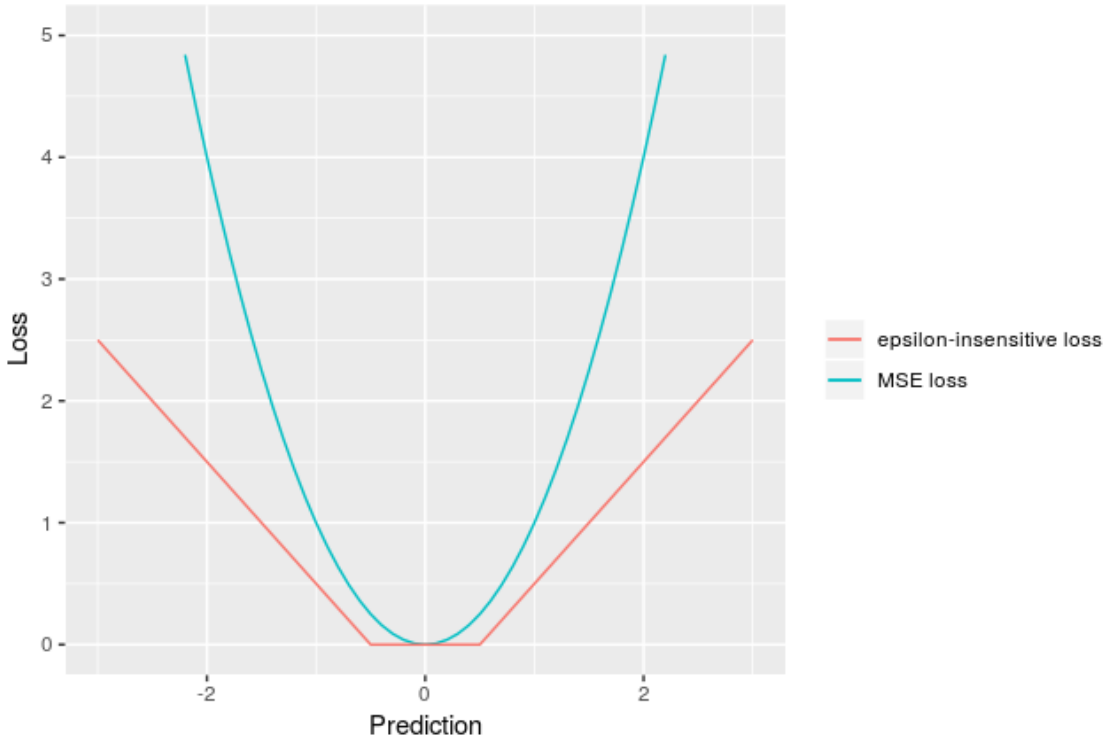


Figure 2.5: Epsilon-insensitive (with $\epsilon = 0.5$) loss function used with ϵ-SVR compared to a MSE loss function (ground-truth = 0). Here it can be seen that errors within $\pm\epsilon$ do not affect the loss function i.e. prediction errors inside $\pm\epsilon$ result in loss = 0.

2.3.1 Support vector regression

SVR is an adaptation by Vapnik [93] of the support vector machine (SVM) classification algorithm [94] for regression purposes. This algorithm seeks to construct decision boundaries that maximise the separation between training data points and a class boundary [94]. The training examples that contribute to the final solution are known as support vectors. As a regression example, linear epsilon-SVR (ϵ-SVR) is discussed in this section. For a prediction value, denoted $\hat{y}$, errors within a range of $\pm\epsilon$ to the gold standard target, denoted y, are tolerated (i.e. not penalised for) using this method, resulting in a soft error margin. The loss function for this regression technique is depicted in Figure 2.5 compared with a mean squared error (MSE), or L2, loss function.

This ϵ-SVR loss function is defined as

$$L_\epsilon(y, \hat{y}) \triangleq \begin{cases} 0 & \text{if } |y - \hat{y}| \leq \epsilon \\ |y - \hat{y}| - \epsilon & \text{otherwise.} \end{cases} \tag{2.1}$$

This results in only points outside $y \pm \epsilon$ contributing to the loss as shown further by the blue points in Figure 2.6.

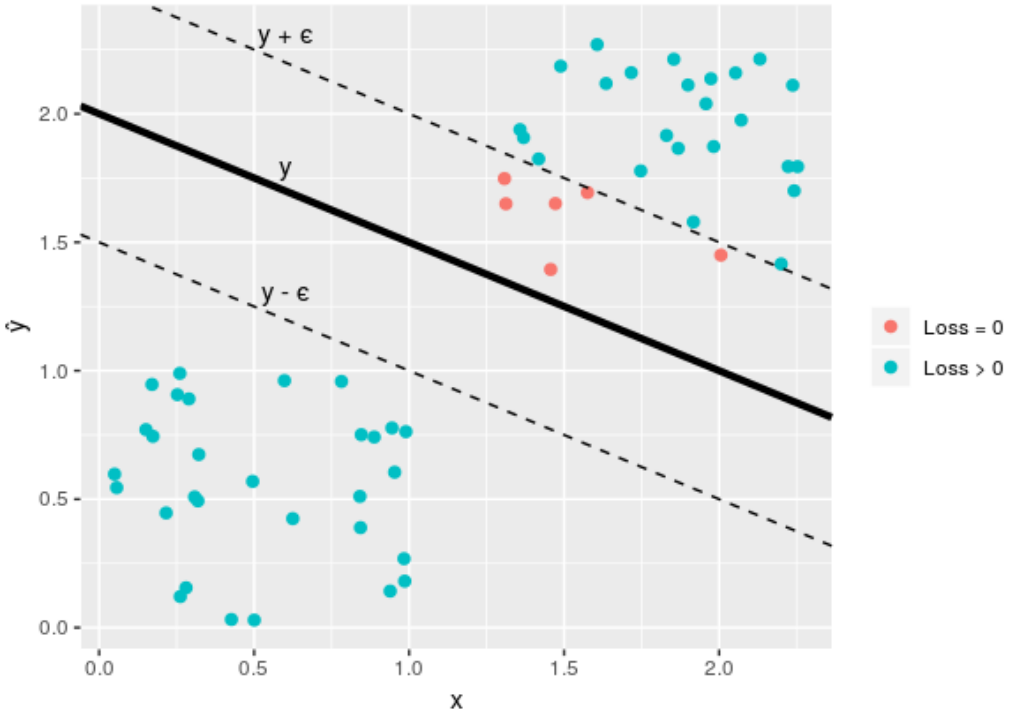


Figure 2.6: Example of predictions which have effect on ϵ-SVR loss function (i.e. $|y - \hat{y}| > \epsilon$) and those which do not ($|y - \hat{y}| \leq \epsilon$). $\epsilon = 0.5$.

Slack variables, ξ^- and ξ^+, are introduced to represent the degree to which errors are greater than ϵ:

$$y_i \leq \hat{y}_i + \epsilon + \xi_i^+ \tag{2.2}$$

$$y_i \geq \hat{y}_i - \epsilon - \xi_i^- \tag{2.3}$$

where $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i + b$ is the prediction, based on the dot product of weights, $\mathbf{w}$, with an input feature vector for a training observation index i, $\mathbf{x}_i$, plus the bias term, b. The objective function J, to be minimised to provide weight estimates, $\hat{\mathbf{w}}$, follows from [93]:

$$\begin{aligned} J &= C \sum_{i=1}^{N} (\xi_i^+ + \xi_i^-) + \frac{1}{2} \|\mathbf{w}\|_2^2 \\ \text{subject to } &\begin{cases} y_i - \mathbf{w}^\top \mathbf{x}_i - b & \leq \epsilon + \xi_i^+ \\ \mathbf{w}^\top \mathbf{x}_i + b - y_i & \leq \epsilon + \xi_i^- \\ \xi_i^+, \xi_i^- & \geq 0 \end{cases} \end{aligned} \tag{2.4}$$

where $C > 0$ is a constant called the complexity parameter and $\|\mathbf{w}\|_2$ is the L2 norm. The C parameter controls the flatness, or, how errors outside $\pm\epsilon$ contribute to the solution. It is equivalent to a $1/\lambda$ regularisation constant and therefore, smaller

C can be thought of as providing higher regularisation. The optimal solution to Equation (2.4) has the form [95]:

$$\hat{\mathbf{w}} = \sum_i \alpha_i \mathbf{x}_i \tag{2.5}$$

where $\alpha_i \geq 0$. This provides a sparse solution called support vectors (i.e. where $\alpha_i > 0$) due to only errors outside $y_i \pm \epsilon$ contributing to the solution. The resulting support vectors for function representation can be very high-dimensional [96].

Predictions in learned models are made using

$$\hat{y} = \hat{b} + \hat{\mathbf{w}}^\top \mathbf{x}, \tag{2.6}$$

while expanding for $\hat{\mathbf{w}}$'s definition gives

$$\hat{y} = \hat{b} + \sum_i \alpha_i \mathbf{x}_i^\top \mathbf{x}. \tag{2.7}$$

With the kernelised (i.e. real-valued function with two arguments [95]) solution:

$$\hat{y} = \hat{b} + \sum_i \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}), \tag{2.8}$$

where in the linear kernel case the original features are used, $\kappa(\mathbf{x}_i, \mathbf{x}) = \mathbf{x}_i^\top \mathbf{x}$.

Advantages of the SVR algorithm in modern day machine learning are the low hyperparameter count and the capability to perform well on small or sparse data sets. By way of example, ϵ and the complexity parameter C are often the only tuned hyperparameters during linear ϵ-SVR. In contrast, deep learning models might have to consider layer count, node count for each layer, learning rate and training iterations among others. SVR helps prevent overfitting by promoting sparsity in solution vectors, while the transformation of features into a higher dimensional space allows the algorithm to prevent underfitting. The algorithm can train models using sparse data quickly due to faster dot product calculations. A disadvantage to SVR, however, includes a potentially unmanageable amount of support vectors with increasing training set size. Also, slower training times may be observed in large data sets due to the way the optimisation problem is traditionally solved (i.e. iterative dual problem optimisation with small steps). However, if researchers can restrict themselves to linear SVR, the tool LIBLINEAR [97] provides a fast solver for SVR. The LIBLINEAR training time can be reduced even further by choosing primal problem optimisation only.

SVR is popular for continuous affect prediction model research [3], [7], [14], [98], [99]. This section has provided an introduction to the SVR algorithm along with discussion of some benefits and drawbacks to SVR. For further details on SVR the

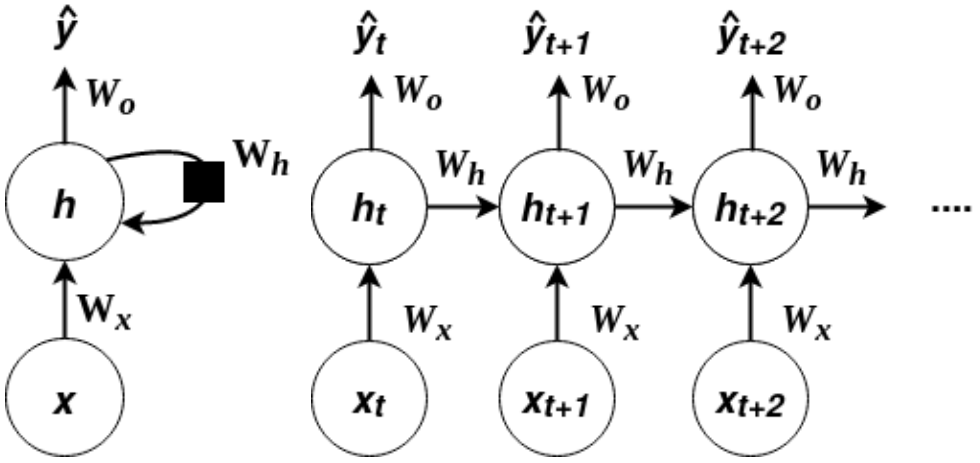


Figure 2.7: A simple RNN illustrated with (a) recurrent connection and, (b) time unrolled graph adapted from [102, p. 369]. W_x, W_h, and W_o are weight parameter matrices for input, hidden, and output layers respectively. Time is indexed with t while the black square indicates a $t-1$ delay.

reader is referred to [96]. In the following section, the most popular dynamic-time regression model algorithm for continuous affect prediction is introduced.

2.3.2 Long short-term memory recurrent neural network

The LSTM-RNN was first presented by Hochreiter and Schmidhuber [100] to avoid the vanishing gradient problem that standard recurrent neural network (RNN) [101] suffer from. These neural networks contain what [100] presented as memory units that allow nodes to store context information during network training. Both RNNs and LSTM-RNNs are intended for use with sequential data, for example, word prediction in written language or electricity demand forecasting.

A simple example RNN with one hidden layer is shown in Figure 2.7 to illustrate the concept of these networks. Here we can see a recurrent connection in the hidden layer, h, where at any $t > 0$ a connection of the $t-1$ to the t layer can be observed. These temporal connections are in addition to weighted input and output connections in standard feed-forward neural networks. The intention is to allow information from previous steps in the sequence to be incorporated in later steps. The shallow network shown in Figure 2.7 (b) depicts output $\hat{y}_t$ at each temporal processing stage t of the network. It is also possible to have only one output at the end of each input sequence or many outputs (a generated sequence) for one provided input.

The RNN considered thus far incorporates past inputs along with the present input for prediction output. Bidirectional RNNs are also possible. Within this

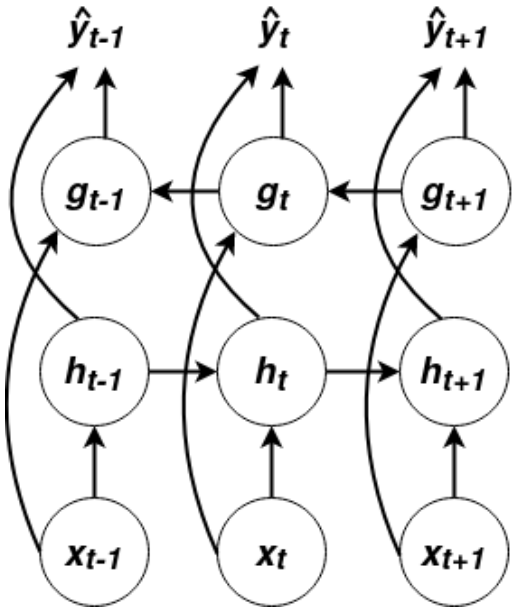


Figure 2.8: Bidirectional RNN adapted from [102, p. 384]. Layer weight matrices omitted for illustration purposes.

variant of RNN, both past and future context can be incorporated to produce output $\hat{y}_t$. Of course, the full input sequence must be received to take full advantage of future context. A simple bidirectional RNN is shown in 2.8.

The weights in the layers of a RNN are generally initialised to some small random values. These values are continually adjusted during model training based on a learning rate and a loss function value until some stopping criteria for training have been achieved. In order to adjust the weights for each connection after a forward pass of the network (illustrated in green in Figure 2.9) the gradient of the loss $L(y, \hat{y})$ is backpropagated through the network. For a RNN, this is called backpropagation through time (BPTT). BPTT involves running the backpropagation algorithm through the network to obtain partial derivatives of the loss with respect to the parameter matrices, e.g.: $\frac{\partial L}{\partial \mathbf{W}_x}$, $\frac{\partial L}{\partial \mathbf{W}_h}$, and $\frac{\partial L}{\partial \mathbf{W}_o}$ for a unidirectional RNN with

$$L(y, \hat{y}) = \sum_{t=1}^{T} L(y_t, \hat{y}_t) \tag{2.9}$$

where $\hat{y}_t = \mathbf{W}_o \mathbf{h}_t$ and $\mathbf{h}_t = \mathbf{W}_x \mathbf{x}_t + \mathbf{W}_h \mathbf{h}_{t-1}$ (nonlinear activations and bias terms omitted for simplicity/illustration as per [103]). The intuition of the BPTT algorithm is shown visually by the red arrows in Figure 2.9. The error gradient can be seen propagating backwards in time through the recurrent layer.

In order to find the derivatives of the aforementioned parameter matrices [103]:

$$\partial_{\mathbf{W}_o} L = \sum_{t=1}^{T} \text{prod}(\partial_{\mathbf{W}_{\hat{y}_t}} L(y_t, \hat{y}_t), \mathbf{h}_t), \tag{2.10}$$

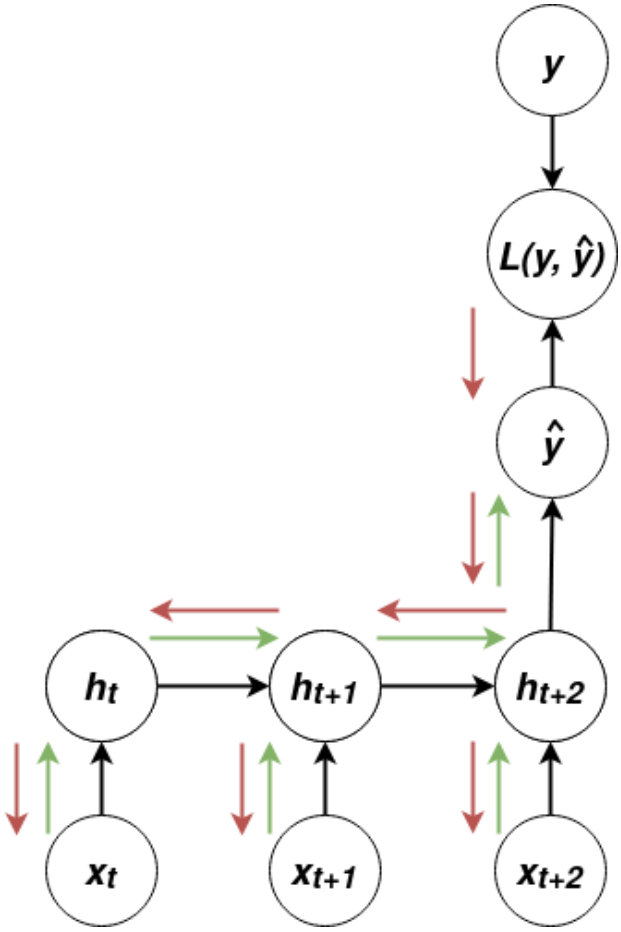


Figure 2.9: RNN forward pass (green lines) and backward error propagation pass or BPTT (red lines) for a RNN with one output $\hat{y}$ per input sequence. The back-propagated error is used for weight parameter value updates.

where prod(.) indicates the product of two or more matrices, provides the output layer derivatives, and

$$\begin{aligned}\partial_{\mathbf{W}_h} L &= \sum_{t=1}^{T} \text{prod}(\partial_{\mathbf{W}_{\hat{y}_t}} L(y_t, \hat{y}_t), \mathbf{W}_o, \partial_{\mathbf{W}_h} \mathbf{h}_t), \\ \partial_{\mathbf{W}_x} L &= \sum_{t=1}^{T} \text{prod}(\partial_{\mathbf{W}_{\hat{y}_t}} L(y_t, \hat{y}_t), \mathbf{W}_o, \partial_{\mathbf{W}_x} \mathbf{h}_t)\end{aligned} \tag{2.11}$$

provides the hidden and input layers' derivatives respectively. Also, the derivatives of the hidden layers with respect to the parameter matrices required for the above calculation are

$$\begin{aligned}\partial_{\mathbf{W}_h} \mathbf{h}_t &= \sum_{j=1}^{t} (\mathbf{W}_h^\top)^{t-j} \mathbf{h}_j, \\ \partial_{\mathbf{W}_x} \mathbf{h}_t &= \sum_{j=1}^{t} (\mathbf{W}_h^\top)^{t-j} \mathbf{x}_j\end{aligned} \tag{2.12}$$

because past hidden layer states affect future hidden states. Within the gradient descent algorithm, for example, the parameter values can then be adjusted with the update rule applied across the weight matrices:

$$\begin{aligned}\mathbf{W}_o &= \mathbf{W}_o - \alpha \partial_{\mathbf{W}_o} L, \\ \mathbf{W}_h &= \mathbf{W}_h - \alpha \partial_{\mathbf{W}_h} L, \\ \mathbf{W}_x &= \mathbf{W}_x - \alpha \partial_{\mathbf{W}_x} L.\end{aligned} \tag{2.13}$$

The constant α above is known as the learning rate and is a real-valued positive number. The learning rate therefore affects how much the derivatives change the parameters at each training step.

An issue with RNNs is that the gradients will tend to vanish, or more rarely, explode with increasing time steps, which results in a lack of parameters training convergence or unstable optimisation respectively. A lack of training convergence was shown for a binary classification RNN trained using gradient descent for input sequences of $T = 20$ time steps [104], for example. Clearly, it would be beneficial to model sequences with more time steps than this, which is what the LSTM-RNN, a gated variant of the RNN intends to provide. A LSTM-RNN cell contains input (or candidate memory), and gated input, output and forget activations as can be seen in Figure 2.10 and the Equations (2.14). The forget gate controls the recurrent cell state connection. Therefore, a node/cell can decide what temporal context to remember and incorporate it into the cell state and the current temporal output.

The following are the forward equations for a shallow unidirectional LSTM-RNN [103] with h hidden units and a batch of training examples of size n with

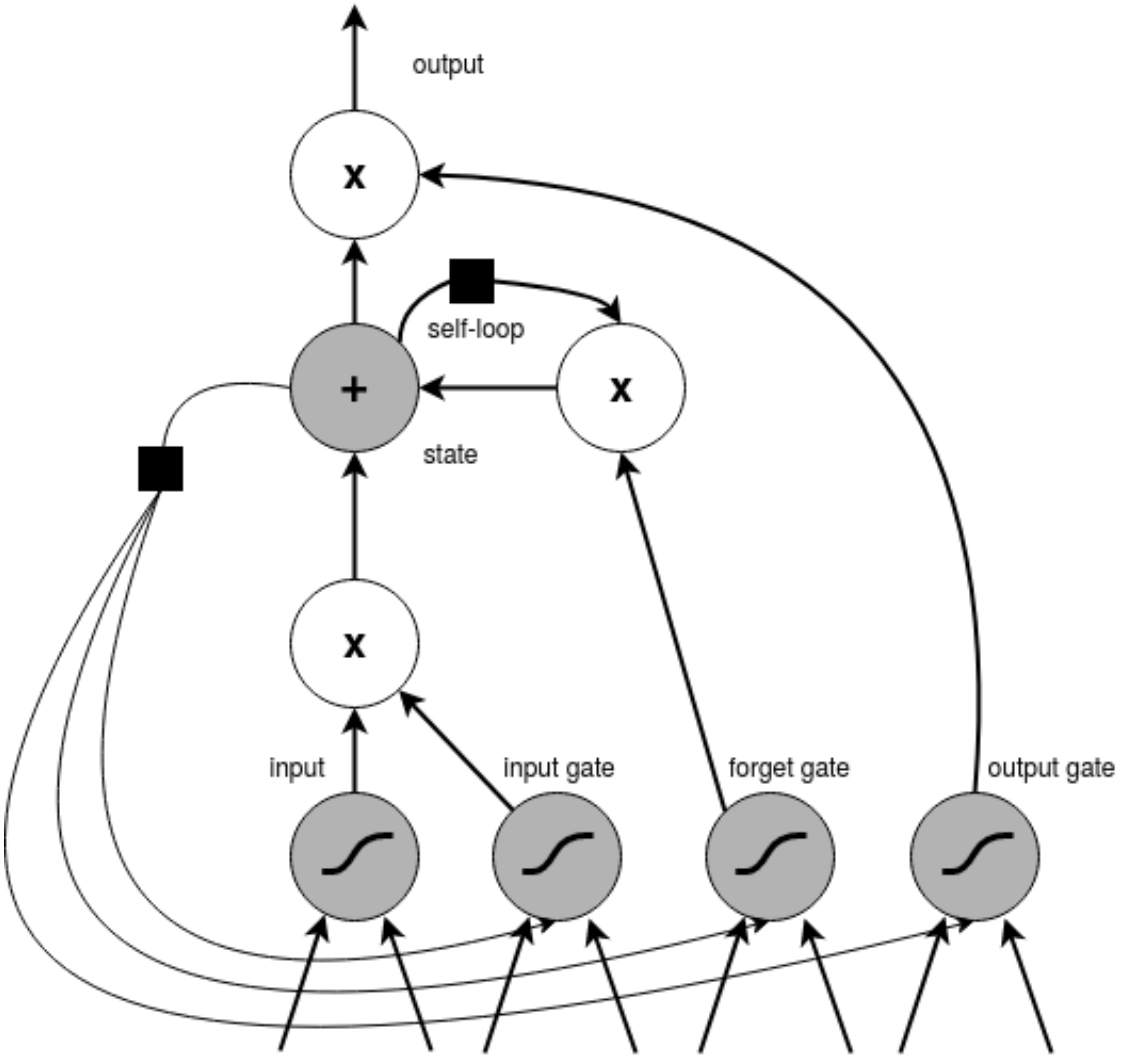


Figure 2.10: Illustration of a LSTM-RNN node/cell adapted from [102, p. 398]. Input and hidden state output at $t-1$ can be seen provided to each of: input, input gate, forget gate, and output gate activations. The cell state, controlled by the forget gate, has a recurrent self-connection. A recurrent connection of the cell state is possible through what are called "peephole" connections to the input, forget, and output gates. Black squares indicate a $t-1$ delay. All multiplications shown are element-wise (also known as Hadamard) products.

dimensionality d and $\mathbf{X}_t \in \mathbb{R}^{n \times d}$ and $\mathbf{H}_{t-1} \in \mathbb{R}^{n \times h}$ (with no peephole connections from the cell state to input, forget and output gates):

$$\begin{aligned} \mathbf{I}_t &= \sigma(\mathbf{X}_t \mathbf{W}_{xi} + \mathbf{H}_{t-1} \mathbf{W}_{hi} + \mathbf{b}_i), \\ \mathbf{F}_t &= \sigma(\mathbf{X}_t \mathbf{W}_{xf} + \mathbf{H}_{t-1} \mathbf{W}_{hf} + \mathbf{b}_f), \\ \mathbf{O}_t &= \sigma(\mathbf{X}_t \mathbf{W}_{xo} + \mathbf{H}_{t-1} \mathbf{W}_{ho} + \mathbf{b}_o), \\ \tilde{\mathbf{C}}_t &= \tanh(\mathbf{X}_t \mathbf{W}_{xc} + \mathbf{H}_{t-1} \mathbf{W}_{hc} + \mathbf{b}_c). \end{aligned} \tag{2.14}$$

$\mathbf{I}_t, \mathbf{F}_t, \mathbf{O}_t \in \mathbb{R}^{n \times h}$, are input, forget, and output gates respectively. Also, $\tilde{\mathbf{C}}_t \in \mathbb{R}^{n \times h}$ is the candidate memory, $\mathbf{W}_{xi}, \mathbf{W}_{xf}, \mathbf{W}_{xo}, \mathbf{W}_{xc} \in \mathbb{R}^{d \times h}$ and $\mathbf{W}_{hi}, \mathbf{W}_{hf}, \mathbf{W}_{ho}, \mathbf{W}_{hc} \in \mathbb{R}^{h \times h}$ are weight parameters and $\mathbf{b}_i, \mathbf{b}_f, \mathbf{b}_o, \mathbf{b}_c \in \mathbb{R}^{1 \times h}$ are bias parameters for each of input, forget, output, and candidate memory respectively additions to input, x, and hidden, h, layer subscripts while σ is the sigmoid function.

The cell state, $\mathbf{C}_t$, is then given by

$$\mathbf{C}_t = \mathbf{F}_t \odot \mathbf{C}_{t-1} + \mathbf{I}_t \odot \tilde{\mathbf{C}}_t \tag{2.15}$$

where $\odot$ is the element-wise product. This equation shows how the LSTM-RNN algorithm forgets as $\mathbf{F}_t$ will take a value between 0 and 1 from the sigmoid activation, controlling how much of memory, cell state $\mathbf{C}_{t-1}$, is retained. Errors can propagate successfully through LSTM-RNNs as only relevant temporal dependencies effectively remain while irrelevant ones are discarded. This is in contrast to standard RNNs with constant error flow through time and thus the vanishing gradient problem appearing with long-term dependency modelling.

Lastly, the hidden state output, $\mathbf{H}_t$, is

$$\mathbf{H}_t = \mathbf{O}_t \odot \tanh(\mathbf{C}_t). \tag{2.16}$$

$\mathbf{H}_t$ can then be used to provide $\hat{y}_t$, where for example one input sequence's hidden activations, $\mathbf{h}_t$, could be connected to a many-to-one fully-connected neural network output layer.

It is also possible to model future and past context together using LSTM-RNNs, where a forward (i.e. past context) LSTM-RNN and a backward (future context) LSTM-RNN are combined into a bidirectional long short-term memory recurrent neural network (BLSTM-RNN). Today, both LSTM-RNN variants are widely used for continuous affect prediction model generation, with good performance achieved in [12] and [13] using deep variants (i.e. more than one hidden layer) of LSTM-RNNs. An advantage of LSTM-RNN and dynamic-time regressors generally, for continuous affect prediction, is the natural fit of generated models for the sequen-

tial, temporal nature of affect. Context (including temporal context) is important for emotion/affect and the temporal component of emotion has been reasoned about since Wundt [50]. Natural temporal components such as *antecedent* [17], *evaluative reactions* [40] and *sequential evaluation checks* [41] have been theorised as part of, comprising, or constructing emotion episodes respectively. The ever-present core affect phenomenon accompanies these emotion episodes or parts thereof and of course can exist outside of emotion episodes. Theoretically, effective modelling of temporal affective changes may be of benefit to continuous affect prediction.

LSTM-RNNs, particularly the deep-layered variants, also provide advantages in terms of feature representation. Deep neural networks combine input feature vectors at each hidden node and subsequent deeper layers. Therefore, feature representations (and combinations of these features) helpful for the task at hand can be learned by adjusting the feature weights. Moreover, in the case of LSTM-RNNs, these feature representations can be learned in a time-dynamic way. Also, from a practical perspective, perhaps due to the popularity of deep learning in modern times, these networks provide practical advantages including freely available flexible programming frameworks and fast implementations of algorithms. These advantages allow the use of various loss functions and gradient-based optimisers, different network topologies/architectures, training strategies and faster experimental and/or hyper-parameter evaluation. Albeit these advantages do come at the cost of increased model complexity, with LSTM-RNNs producing, arguably, some of the most complex and least interpretable of today's available machine learning models.

This section has provided an introduction to the RNN and LSTM-RNN algorithms, providing information on the algorithms' operations and theoretical advantages for continuous affect prediction. For further details on these learning algorithms and related items (e.g. vanishing gradient, optimisation) to RNNs the reader is referred to [102] and [103]. The following section introduces the predominant performance evaluation measure currently used for continuous affect prediction.

2.3.3 Concordance correlation coefficient

Continuous affect prediction cannot rely on precision, recall or accuracy measures based on true/false positive and true/false negative measures associated with classification due to the required continuous-valued regression/prediction problem to be solved. The CCC value, however, was proposed by Lin [105] as a measure of agreement and a reproducibility suitable for continuous-valued measurements. It can take on values ranging $[-1, 1]$ and provides a measure of both precision (trend following) and accuracy (error) to overcome issues associated with other measures of continuous-valued reproducibility. Examples of these problematic measures for

reproducibility include Pearson's correlation coefficient, r, which measures precision only and least squares, which can fail under very small or large errors [105]. Weninger et al. [106] graphically illustrated issues with using Pearson's r as a continuous affect prediction evaluation metric where they showed its invariance to scaling and shifting compared to CCC. A measure that considers the trend, but not the deviation from gold standard values can be problematic for affect prediction as predictions are not penalised for under- or over-shooting the true gold standard value. The formula for Pearson's r, for comparison to CCC is defined as:

$$r = \frac{\sigma_{\mathbf{y}\hat{\mathbf{y}}}}{\sigma_{\mathbf{y}}\sigma_{\hat{\mathbf{y}}}}, \tag{2.17}$$

where $\mathbf{y}$ are ground-truth (or gold standard, in the case of affect prediction) values, $\hat{\mathbf{y}}$ are predictions and $\sigma_{\mathbf{y}\hat{\mathbf{y}}}$ is the covariance. An example of the difference between r and CCC is given is Figure 2.11 for a small paired-data sample. In the figure caption it can be seen that the perfectly correlated values y and $\hat{y}$ produce very different r and CCC values. The large difference is due to the easily observed squared-error, or shift of -0.5, between y and $\hat{y}$ that CCC is sensitive to. The CCC for a paired-data sample is given by

$$\text{CCC} = \frac{2\sigma_{\mathbf{y}\hat{\mathbf{y}}}}{\sigma_{\mathbf{y}}^2 + \sigma_{\hat{\mathbf{y}}}^2 + (\mu_{\mathbf{y}} - \mu_{\hat{\mathbf{y}}})^2}, \tag{2.18}$$

where σ^2 denotes the uncorrected sample variance and μ is the mean.

The CCC is an important performance measure for continuous affect prediction, where both precision and accuracy measures are desirable for performance estimation. The CCC allows effective measurement of both of these performance metrics in one term. This performance measure is largely used today in continuous affect prediction research to assess model performance. Moreover, researchers are starting to use CCC as part of their loss function for model training with improved performance over MSE in some cases, for example [33], [106]. An advantage of CCC over MSE promoted by [106] is that training targets need not be standardised when using this metric as a loss function. However, they also note that a MSE loss function does maximise correlation while minimising error for standardised target values.

Of course, the linear correlation relationship specified by r values are useful in affective computing, for example, where error measures are inappropriate such as data naturally measured on different scales. An example of this is input feature to arousal or valence target linear relationship assessment. Ideally features would score high in terms of their linear relationship with either dimension of affect, indicating that they may be predictive of the respective target affect.

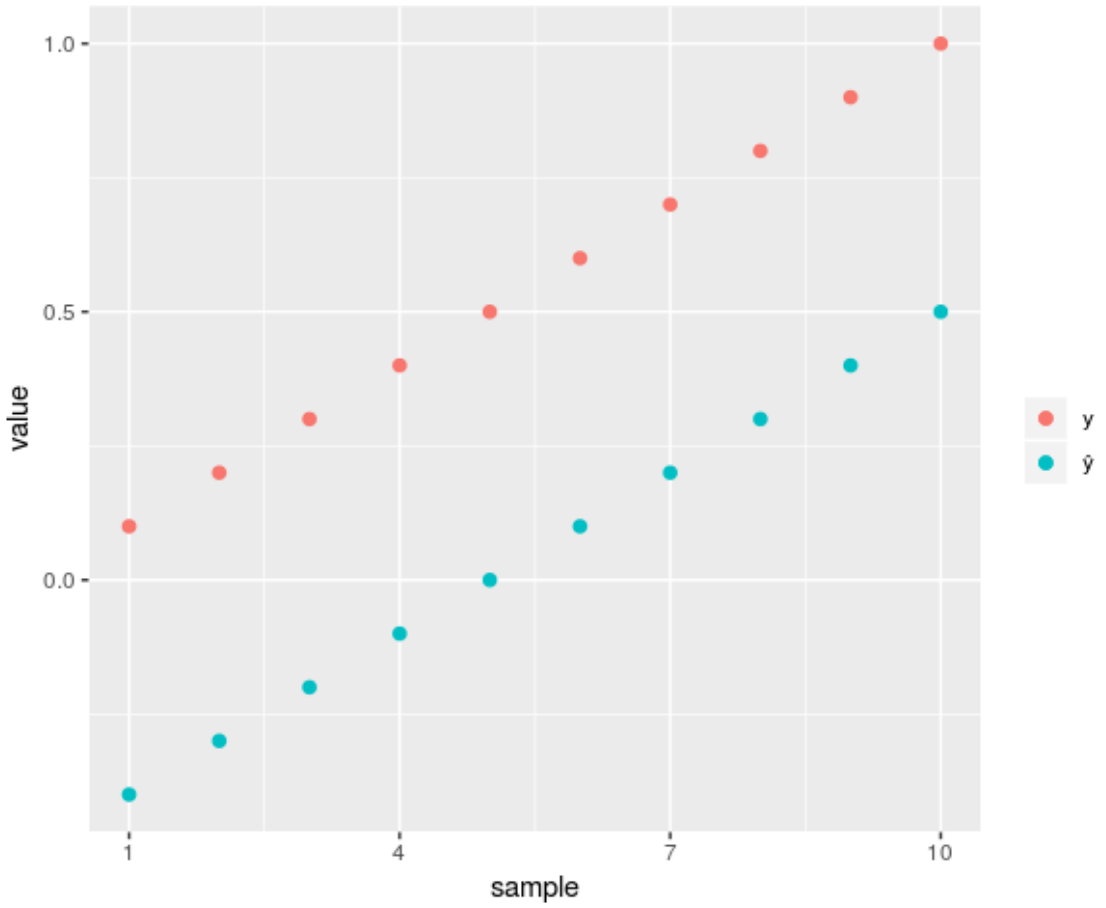


Figure 2.11: Visualisation of ground-truth, $y = \{0.1, 0.2, ..., 1.0\}$, and predictions, $\hat{y} = \{-0.4, -0.3, ..., 0.5\}$. Estimates of r and CCC are 1 and 0.398 for the sample respectively.

2.3.4 Affect learning and prediction

Learning and predicting continuous affect is challenging, prompting increased efforts in modelling affective phenomena using machine learning algorithms. Some challenges for modelling include obtaining effective input features, addressing temporal delays in gold standard annotations, effectively leveraging obtained annotation values during modelling and feature stream fusion (for multimodal systems). This section explores what efforts have been undertaken to address these challenges by the continuous affect prediction research community. This section of the literature review focuses solely on speech-, head- and eye-based continuous affect prediction. The interested reader is directed to the survey by D'Mello and Kory [107], and the review by Osman and Falk [108], for further information on modalities and affective computing tasks outside this focus of this review.

2.3.4.1 Speech features

Numerous handcrafted feature sets are available for continuous affect prediction. These sets have been developed by experts in paralinguistics and computer science and there is ample evidence to suggest that speech-based cues convey the various affective states of speakers [109]. Ringeval et al. [12] and Mencattini et al. [110] used 65 LLDs based on the Interspeech Computational Paralinguistics ChallengE (ComParE) 2013 speech feature set [31] and their first-order derivatives as speech input for continuous arousal and valence regression. The LLDs in the Interspeech ComParE 2013 feature set [31] consist of measures such as loudness, RMS energy, zero-crossing rate (ZCR), Mel-frequency cepstral coefficients (MFCC) [111] (MFCCs 1-14), fundamental frequency (F0)-based, and voice jitter and shimmer measures. The LLDs are gathered from prosody, spectral, cepstral, and voice quality speech feature groups, which are known to perform well for speech processing, music information retrieval and voice-based pathology analysis tasks [112]. These LLD features, and the suprasegmental features gathered based on the LLDs (functionals of the LLDs), are discussed in-depth Weninger et al. [112]. Test set CCC results achieved in the literature using the ComParE LLDs, where min, max, range, mean and SD functionals were applied to the LLDs, are 0.804 for arousal and 0.528 for valence as part of multimodal input to LSTM-RNN-based networks [12].

Perhaps the most popular handcrafted speech feature set for continuous affect prediction is the extended Geneva minimalistic acoustic parameter set (eGeMAPS) set [113] developed by Eyben and colleagues. These features have been included in numerous audio-visual emotion challenge (AVEC) events as baseline features [3], [8], [9], [90], [114]. The base set, Geneva minimalistic acoustic parameter set (GeMAPS), and the extended version, eGeMAPS, are proposals for shared standard feature

sets for use in voice research and affective computing. These feature sets have a relatively small size compared to other sets. eGeMAPS with a dimensionality of 88, for example, is favourable in terms of size compared to other proposed sets such as ComParE 2013 [31] and AVEC 2014 [7] with dimensionalities of 6,373 and 2,268 respectively. Compared to ComParE 2013 [31], the eGeMAPS set includes vocal tract (formant frequency features), harmonic difference measures and a smaller number of MFCC (i.e. MFCCs 1-4) LLDs. The justification for the inclusion of the former LLDs includes evidence of association with emotion, cognitive load, and pathology. The latter (small number of MFCCs) LLDs are argued to be the most important MFCCs for affect and paralinguistic analysis. Various means, standard deviations, rates, count and percentile functionals are then gathered for a specified time-window and rate to complete the feature set for GeMAPS/eGeMAPS. Some practical advantages of eGeMAPS include increased ease of model interpretability and a lower model training time compared to larger feature sets. Brady et al. [14] achieved test set CCC values of 0.770 (arousal) and 0.687 (valence) for their system, which included eGeMAPS and other audio features as part of their multimodal submission.

Phoneme-based features have also been investigated for speech-based continuous affect prediction. These are short and perceptually distinct sounds that, when combined with other phonemes in a sequence, provide linguistic content from speech. Huang and Epps [115] investigated phone log-likliehood ratio (PLLR) features for continuous affect prediction, and they proposed a phonetically-aware acoustic feature set from speech. Each PLLR feature in [115] consisted of taking the log of the probability of a phoneme in question divided by the average probability of all the other phonemes in the set. Each phoneme in the set was modelled by a respective hidden Markov model (HMM). When compared against traditional speech features, the PLLR features performed comparably or better for arousal and valence prediction, with more improvement observed for valence than arousal. The authors also proposed a phonetically-aware acoustic feature set, PA-eGeMAPS, by repeating the 25-dimensional eGeMAPS LLD features for each of 39-dimensional English phonemes. The proposed 975-dimensional feature set consisted of functionals (mean of each LLD) that were weighted based on the posterior probabilities of the phonemes and a weighting parameter that controlled the phonetic-awareness of the features. CCC results achieved using PA-eGeMAPS features evaluated using leave-one-subject-out cross-validation (LOSO-CV) ranged from 0.426 to 0.735 for arousal and 0.221 to 0.429 for valence across a range of corpora tests. Phonetic weighting parameters used ranged from 0.05 to 0.2 indicating phonetic-awareness can benefit continuous affect prediction in speech, however, a larger overall feature vector was used compared with eGeMAPS, for example.

Bag-of-words (BoW) representations of audio signals have been successfully used for continuous affect prediction [99], [116]. The bag-of-words representation, originating from natural language processing, is a word histogram that is generated from the frequencies of each word in a sequence from a dictionary that is searched. For example, a bag-of-audio-words can be obtained using unsupervised learning by taking word representations from a codebook of user-specified size based on LLDs such as MFCCs. An exemplary test set CCC performance for arousal prediction using bag-of-audio-words is 0.753 by Schmitt et al. [99], based on SVR on the REmote COLlaborative and Affective interactions (RECOLA) corpus [16]. Of note in [99] was the larger temporal window required for feature extraction for valence prediction compared to arousal for optimal performance. This difference in temporal window for different affects, and modalities, was also found by Valstar et al. [3]. BoW-based feature representations are also possible in text and video. BoW-based features have been provided as baseline feature sets for AVEC since 2017 [9], [90], [114].

A trend that has developed in speech feature extraction for input to learning algorithms is deep learning of features. Using deep learning, features can be learned in a supervised (e.g. end-to-end paradigm) or unsupervised way. In either case, feature representations are learned from the given data (e.g. raw, minimally processed audio, or LLDs) before being used as input to an algorithm for affect prediction. This enables automatic learning of useful representations of the data and successful prediction with less feature engineering effort. In an end-to-end framework, concatenation of feature extraction and affect prediction networks is required. Deep learned representations of speech features for continuous affect prediction have been used on their own [33], [34], [117]–[119], and combined with traditional handcrafted features in [120], [121]. Trigeorgis et al. [33] used a combined convolutional neural network (CNN)-to-BLSTM-RNN architecture in which 1-dimensional temporal convolution kernels were learned for feature representations from raw audio and were applied as

$$(\mathbf{f} * \mathbf{x})(t) = \sum_{k=-T}^{T} f(t)x(t-k) \tag{2.19}$$

where $\mathbf{f}$ is the kernel to be convolved with $\mathbf{x}$. Feature transformations, after max pooling (taking maximum feature values for a specified segment) and perhaps further convolutional and pooling layers, are then passed to a BLSTM-RNN for affect prediction. Chen et al. [120] used features from SoundNet [122], a novel convolutional sound network trained using state-of-the-art image processing model labels, as part of their multimodal submission to AVEC 2017 [90]. They achieved an arousal CCC of 0.672 and a valence CCC of 0.756 for continuous prediction on the test set, showing good generalisation capability for a system considering these features.

What is further interesting in the approach taken in [120] is that they consider the interlocutor's audio features in the continuous affect prediction of subjects, effectively multiplying the input feature vector by 2. They achieved better performance with this approach than mixed audio (both interlocutor and target subject) features together for prediction or zeroing out the interlocutor's audio features and considering the subject's speech on its own for prediction. Subsequently, Huang et al. emulated this approach with success [121]. Considering the interlocutor's influence (i.e. features), when data such as speaker turn-taking is available, appears to be a good strategy toward more effective continuous affect prediction in dyadic scenarios based on these results.

2.3.4.2 Head and eye features

Features from the visual domain can be extracted using 2-dimensional convolutions [34], dynamic space-time appearance (visual texture) descriptors [123] or geometric (measurement/tracking-based) approaches [124]. These approaches to feature extraction from face images, for example, implicitly contain head- and eye-based information, however, the contribution of the head and eye features to the overall visual affect display is not clear from these features. Furthermore, instead of pixel intensities or textures, higher-level features in the form of specific head- or eye-based gestures from the visual domain could provide interpretability benefits for affective computing systems. For example, downward head movement was associated with sadness in [26] while some specific facial actions (or combinations of facial actions) are often associated with certain visual displays of affect [18].

Gunes and Pantic [72] used HMMs to learn temporal head-based cues (nods, shakes, other head movements) based on directional codeword cues obtained from video. The directional codewords included rightward, upward, leftward, and downward head motion. The total feature vector included: total duration of head movement (codewords), mean and SD of head movement angle and magnitude, log likelihoods outputted by the HMMs, and a maximum likelihood classification vs the HMM outputs. SVR was used for continuous affect learning and prediction based on these input features. Results comparable to an independent speech-based system of the time were achieved. Eyben et al. [125] used a similar feature vector to [72], however, they also used event-based head nod, head shake, and other head movement features based on the HMM outputs in their evaluations. That is, the head-based events were present (equal to 1) or absent (equal to 0) in an utterance and these features were combined with Ekman's facial action coding system [18] action unit (AU) feature events to complete the authors' visual event-based features. Both functionals of signal features and event-based audio features were also investigated

in [125]. The event-based features were combined into a binary string for BoW-based feature representation. SVR prediction carried out showed the multimodal event-based features to perform best for valence prediction, while the best result for arousal used either audio or audio and visual signals combined with multimodal event-based features. The results in terms of r were 0.699 for arousal and 0.165 for valence and these scores did not surpass group-of-humans correlation performance estimates of 0.704 and 0.818 respectively.

Head pose was used with AU features for visual feature input to multimodal (speech, visual, physiological features) continuous affect prediction systems by Ringeval et al. [12]. The head pose features included three-dimensional static pose and two short-term dynamic features, the mean and SD of the optical flow region around the head. Visual features outperformed audio features for valence prediction while the reverse was true for arousal in the experiments. The best multimodal affect prediction systems always included both the audio and visual features. CCC scores of 0.804 for arousal and 0.528 for valence were achieved using SVR-based model fusion of LSTM-RNN models on the RECOLA [16] test set.

More recently, Wu et al. [88] used head pose and eye gaze cues gathered from video to (1) guide an attention mechanism for learning features from face images and (2) augment facial features for continuous affect prediction. The attention mechanism used softmax probabilistic output for the weighting of facial features, so that features in the sequence were given less weighting in the presence of extreme head rotation, for example. Furthermore, the authors of [88] combined a high-level CNN-learned pose and gaze feature of dimensionality 1 with the facial features to provide additional information for prediction. The attention and/or combined pose & gaze feature augmentation block improved performance compared to prediction based on the face alone (i.e. the baseline). Unbiased validation set CCC results (i.e. training or hyperparameter experimentation did not use the validation set) were 0.603 for arousal and 0.686 for valence on the RECOLA [16] corpus. Relative CCC performance increases of 15.05% and 12.64% above baseline scores were obtained for arousal and valence respectively. This work showed that head pose and eye gaze can contribute positively to continuous affect prediction in video.

2.3.4.3 Annotations delay compensation

Weighting annotator ratings, based on the correlation of their ratings with the average cohort ratings [126] or how informative a sample of training annotations are [119] has been investigated to improve gold standard quality [3], [9], [90], [114], [119]. It is also accepted by the community that annotators do not instantly provide their ratings; there is a time delay between observation instant and annotator input.

BLSTM-RNN/LSTM-RNNs have been used to address this issue [12] as it is implicit in using these algorithms, designed for sequence modelling, that the network can learn temporally salient features and delays. Improved performance was obtained for the LSTM-RNNs compared to feed-forward neural networks for arousal and valence prediction, which are unable to take advantage of temporal context in [12]. Also, explicitly altering the time at which input features and gold standard targets are aligned prior to model training has resulted in improved LSTM-RNN performance [127]. Some researchers shifted targets prior to training LSTM-RNNs irrespective of their dynamic modelling capability [13], [36]. For example, He et al., altered the temporal alignment of features based on the average CCC between input feature vectors and training targets before feature selection and BLSTM-RNN training [13]. They showed a clear difference in average feature-to-target CCC by altering temporal offsets between inputs and training targets. Others have shown performance increases on validation set SVR by altering temporal offsets similarly prior to model training [3], [90], [99]. Mencattini et al. [110], the authors go one step further and estimate annotator reaction lags based, not only on individual affect dimensions, but also on whether the affect is positive or negative. In their unique approach, which they call quadrant-based temporal division, annotator reaction lags and feature vectors are estimated based on maximising features-with-target correlation using correlation-based feature selection after each temporal shift.

Khorram et al. [128] developed a new CNN node that was designed for a multi-delay sinc network, which comprises both standard CNN nodes and delay-provider nodes specifically designed to learn delays between input features and targets. The delayed sinc kernel used in multi-delay sinc network was implemented with a windowed sinc function to approximate the Dirac delta function to provide temporal delays by lowpass filtering. The bandwidth of the sinc kernel and rectangular temporal window size were pre-selected for these nodes while the time delay parameters were learned. Results obtained for this network, using 40 mel-frequency bank input features, did not outperform implicit annotator delay compensation (i.e. LSTM-RNN [12]) for audio only arousal prediction on the RECOLA [16] test set, but did so for valence. Also, the results obtained on the German-language subset of SEWA [53], outperformed LSTM-RNN [120] (on validation set) but they could not outperform the explicit annotator reaction lag compensation used with CNN in [129]. Ouyang et al. [130] used an auto-regressive model with exogenous variables for speech affect prediction. This model can use time-delayed versions of $\hat{y}_t$ and $\mathbf{x}_t$ while a delay between features and target variables is also possible in addition to current $\mathbf{x}$ values for its prediction at time t. The authors obtained competitive audio only results on the AVEC 2016 [3] and 2017 [90] development sets, but they did not provide test set results so model generalisability cannot be fairly assessed.

2.3.4.4 Feature selection

When generating models for continuous affect prediction, it can arise that not all input features in a feature set are relevant to the task at hand. Further, there may be redundant features that can affect model performance, or there may be simply too many features present in a set for an algorithm to generalise well on unseen data. Also, with large input feature vectors, algorithms will take longer to train models and model interpretability enters a larger (i.e. dense) space. This can result in less progress for both model development and model interpretation efforts. In summary, the goals of feature selection are to (1) reduce the number of input features in the feature space, and (2) improve model performance. In order to achieve these goals, researchers apply a feature selection algorithm to reduce the feature space and/or select the good subset of features for the task at hand. Optimal feature selection is a very difficult problem in computer science, it is non-deterministic polynomial-time hard, and therefore algorithms often use heuristic methods in order to select some reasonable subset of features.

Affect prediction algorithms have been used that incorporate feature selection automatically, such as L1 regularised regression [130] which can effectively "zero-out" features and encourage sparsity due to the L1 penalty applied. In further supervised approaches, correlation-based feature selection has been used [13], [110], [131], which selects features that have higher feature-with-target correlation and lower feature-with-feature correlation to maximise feature relevance and minimise feature redundancy. A similar approach to feature relevance can also be achieved using mutual information (MI) estimation instead of correlation such as the mRMR algorithm [132] which was shown to improve affect prediction performance by Paul et al. [133]. Further heuristic algorithms include sequential forward selection (SFS), which iteratively adds features one by one until no performance improvement is observed for some number of iterations, and filter-based methods. With a filter, a simple feature-with-target association measure can be applied and features outside a defined threshold removed. Principal component analysis (PCA) has been used for unsupervised feature selection [13], [121] and cross-cultural feature adaptation [134] in continuous affect prediction. PCA finds orthogonal projections in data such that the projection error is minimised and data are represented as a subspace of the original data. Users can then select the first k principal components or base their selection of k components on retaining some percentage of the original data's variance to achieve dimensionality reduction compared to the original features. An advantage to PCA is that it can be done on test set or unseen data [134], with the limitation of course that the whole test set must be provided in advance. However, PCA transforms the original features to achieve dimensionality reduction, which

may be undesirable if strict feature reduction is required. Amiriparian et al. [131] provided an evaluation of different feature selection algorithms for continuous affect prediction. Among the numerous techniques employed, SFS was used to achieve the best arousal prediction performance on both validation and test sets. SFS also performed best for valence prediction on the validation set, while feature selection was not able to improve any test set evaluations for this dimension. The authors concluded that valence requires more features than arousal for prediction, thus rendering their feature reduction approaches ineffective.

2.3.4.5 Affect modelling algorithms

Continuous affect prediction work by Nicolaou et al. [135] showed BLSTM-RNN to outperform SVR by way of root mean squared error and Pearson's r performance metrics. They argued that BLSTM-RNN is particularly suitable for continuous affect prediction due to the temporal nature of affective expression features (onset, apex, offset). Today, LSTM-RNN/BLSTM-RNNs are by far the most popular of learning algorithms for continuous affect prediction model generation [13], [33], [34], [36], [119]–[121], [127], [136]. However, emotion episodes may be short-lived, with Ekman observing that the majority of emotions exist in the 0.5 to 4 seconds range [137]. Moreover, it is not always clear what sequence length authors use as input to their LSTM-RNN/BLSTM-RNN models [36], [136], much less what amount of context is incorporated by humans (already cognitively loaded) in providing continuous annotations.

Within deep learning approaches, researchers have further used joint-learning in the form of multi-task learning (MTL) [12], [120], [138] to effectively leverage affect dimension correlations during modelling. MTL involves simultaneous learning of prediction tasks in one network, which for continuous affect prediction could include arousal, valence and dominance learning, for example. In this way, the overall loss to be used for backpropagation can be the sum or weighted sum, of individual (e.g. arousal, valence) losses in the hope of capturing feature weights which better describe affect. MTL has been used to varying degrees of success, with Chen et al. [120] and Parthasarathy and Busso [138] observing increased continuous affect prediction performance for MTL compared with single-task learning, although the converse has also been shown [12]. In [12], MTL was superior for valence prediction using audio features and arousal prediction using video features. However, the authors also showed single-task learning to be superior for arousal prediction using audio and valence prediction using video. This experiment and the trend in the community to model individual affect dimensions separately [13], [14], [33], [36], [121], [127], [128], [136], [139], perhaps indicates that MTL may not always be the best option in terms

of performance.

Sridhar et al. [139] demonstrated that higher regularisation is required for more effective valence prediction from speech when using a deep feed-forward neural network (DNN). They say that valence cues in speech are more speaker-dependent. Improved performance using higher dropout, where nodes are randomly not used for model learning, with probability 0.7, was observed for valence in their work while arousal and dominance networks performed best with a node-use probability of 0.5. Another type of MTL that has been carried out in affect prediction, is modelling individual annotator's ratings for one single affect dimension. Ringeval et al. [12] took this approach and found that valence prediction can be improved when modelling all 6 annotation streams instead of just the mean of all 6 raters. This result did not hold for arousal prediction from speech, however, so perhaps individual annotation modelling better captures the idiosyncrasies of valence.

Schmitt et al. [129] used a CNN to investigate the need for recurrence for continuous affect prediction. The authors showed improved performance on the SEWA [53] German test partition compared to their previous work [140] for arousal prediction using LSTM-RNN, however, they did not do so for valence. Their results are comparable to, and in some cases better than, related work using LSTM-RNN. Perhaps this was because CNN can learn local temporal and spatial features and these features might be more relevant to the task, albeit further empirical findings are required to support this.

Outside of deep learning, SVR remains popular for continuous affect prediction, where it has been used as the main algorithm [131], as part of a multiple regression system [14], [136] and as a fusion regressor [12], [127], [136]. Murphy [95, p. 506], however, argues that there are better alternatives. The affective computing community has investigated some alternatives that Murphy suggests, and others. Relevance vector machine (RVM), a sparse solution technique that provides probabilistic output, has been used along with Gaussian mixture regression (GMR) and Gaussian process regression (GPR) for continuous affect prediction by Dang et al. [141]. The authors of [141] used GMR and GPR to model the affect (mean of the Gaussian) and uncertainty (SD of the Gaussian) due to the inherent ambiguity in the affect ratings provided by annotators. The authors used the dominant mixture, the Gaussian from the mixture group that maximised the conditional probability of targets, given the features [142], to provide their GMR output. GPR, a kernel-based regression method, provides probabilistic output by modelling function probabilities given some data and an additive kernel was used to model temporal feature relationships. In this way the authors modelled the joint probability of input features and affect targets (GMR) and the temporal dynamic of the input features (GPR). Fusion of these two systems with a RVM system was carried out using a further

RVM for the final regression. GPR was not shown to benefit arousal prediction but did improve valence prediction which perhaps suggests that the temporal dynamic of features is important for valence prediction.

Partial least squares (PLS) regression was used to create single-speaker regression model (SSRM)s for affect prediction, where the SSRMs act as input to a cooperative regression model (CRM) for speaker affect prediction by Mencattini et al. [110]. The CRM averages the SSRM predictions dynamically for a temporal window based on consensus, which is the average CCC of a SSRM with all other SSRMs that fall into the 60^{th}-percentile or under. For comparison, the authors also investigated SVR for SSRM building. The results showed that the SVR performed better than the PLS regression method for SSRMs. However, the PLS method performed best for building the CRM, which was the more general predictor overall. The authors suggest that SVR is more prone to overfitting and simpler regression techniques such as PLS may be more suitable for ensemble methods similar to the CRM method presented by the authors.

Huang et al. [136] fused BLSTM-RNN, SVR, and PLS predictions by using another SVR on provided predictions. They achieved novel performance on the RECOLA [16] corpus validation set but this data partition was used in training and parameter evaluation and it is therefore difficult to assess model generalisability from their work. Han et al. [15] also fused multiple regression techniques in their unique Strength Modelling approach to continuous affect prediction. Their technique is so called as they aimed to leverage the strengths of both SVR (global optimum solving) and BLSTM-RNN (temporal context modelling) while minimising each algorithm's respective weaknesses. Strength Modelling uses predictions from an initial model combined with input features as input to a final prediction model. The results in [15] showed that this approach can improve affect prediction performance.

Regularised linear regression (LR) techniques have also been used for affect prediction [130], [143]. Huang and Epps [143] used regularised LR for learning state and observation transition matrices for Kalman filter-based continuous affect prediction from speech. The Kalman filter is a linear-Gaussian state space model that can provide probabilistic output (mean and covariance) and it is composed of state and observation models with Gaussian noise assumed for each. The Kalman gain controls how much errors affect the observation model which is in turn used to provide $\hat{y}$ predictions based on the state model's output plus the Kalman gain weighted error, i.e.

$$\hat{y}_t = \bar{x}_t + G_t(y_t - \bar{x}_t), \tag{2.20}$$

where G_t is the Kalman gain, $\bar{x}_t$ is the initial prediction and y_t is an observation

model output. The authors used a Kalman filter for continuous affect prediction based on arousal, valence, and first and second-order differences of these dimensions. They fused this approach with RVM and used speech-based input. They also used a Kalman filter strictly for RVM output fusion where they improved upon the AVEC 2016 baseline [3] audio valence and arousal validation set scores.

2.3.4.6 Multimodal fusion

The problem of fusing multiple input feature streams for continuous affect prediction is motivated by the potential for improved model performance. For example, for a speech cue at a given time, perhaps a downward head tilt might indicate sorrow visually. It is this complimentarity that researchers seek when fusing multiple feature streams for affect prediction. Primarily, fusion schemes that have been investigated fall into the categories: feature, model and decision fusion [144]. For feature fusion, researchers concatenate signal-based features (i.e. functionals) prior to modelling the required affect(s) and a larger input feature vector to the model results. A variation on feature fusion, called string-based fusion, also combines features prior to modelling except in this case the feature vectors are based on the fusion of event-based presence/absence binary features [125]. Advantages to these approaches are that fewer models need to be trained and, arguably, a simpler, more interpretable model can result. Furthermore, cross-modal feature interactions can be learned in these models. A disadvantage to these approaches, however, is that feature streams may require synchronisation prior to modelling. LLD features occurring at different sampling rates may have to be downsampled prior to feature extraction, invariably introducing information loss.

Model and decision fusion involves integrating trained model predictions from each modality model to form one final prediction for the ensemble. Decision fusion uses a voting or weighting strategy for this, where a model can be used to learn the decision weights. For example, SVR [9], [12], [114], [121], [127], [136], LR [3], [8], [131], and Kalman filters [14], [143] have been used to this end. Model fusion treats modality-wise predictions as intermediate representations, then, as part of the overall learning process, the representations are combined to form input for (further) latent space modelling and final prediction. LSTM-RNNs have been used for this latent space modelling of the intermediate modality representations [13], [118], [135]. An interesting model fusion technique proposed by Nicolaou et al. [135] is output-associative (OA) model fusion. For OA fusion, both arousal and valence predictions from models are used for the final arousal or valence prediction. The goal is not only to provide a prediction based on multiple input modalities but also to leverage output correlations in determining the final prediction. Some advantages to model

and decision fusion approaches include the relaxation of feature synchronisation requirements and the potential to better model individual asynchronous feature stream information. A clear disadvantage to these approaches is that there are more models to train, resulting in a more complex overall model to manage (an ensemble of models).

Chen et al. [118] combined both early feature fusion and model fusion to achieve 2^{nd} place in the AVEC 2019 challenge [114]. In their approach they combined deep learned audio and video feature streams prior to modelling using BLSTM-RNN, and they also trained models using the individual modalities. Then, all the trained models were combined using a further BLSTM-RNN that used the individual and early-fused feature stream predictions as input. Another novel fusion approach that has been researched by Han et al. [144] is implicit fusion, which was used for speech- and face-based continuous affect prediction. Implicit fusion shares some similarities with model fusion as unimodal model layers are later combined in multimodal shared layers while it is similar to MTL in that multiple modality losses are optimised during training. In [144], unimodal models, one for each modality, were learned by using both audio and video data in training where the auxiliary modality was weighted to contribute to a joint-learning loss. For prediction, only one modality was used in their experiments. Implicit fusion provides an advantage of enabling multiple modalities to be leveraged in model training while facilitating prediction where a modality is missing from an audio-video sequence.

Feature and decision fusion methods have been compared in the literature. Ringeval et al. found the best performance with decision fusion in [12] while Amiriparian et al. [131] achieved the best performance with feature fusion. In comparing these, better audio arousal prediction performance was achieved in [12] while better valence performance was obtained by [131] on the RECOLA [16] corpus. It must be noted that different algorithms and features sets were used in [12] and [131]. On the SEWA [53] corpus, the comparison work is not much clearer. In [121], the authors found that feature fusion performed better than decision fusion for 2 of 3 validation set evaluations. For their test set evaluations, feature fusion only performed better for 3 out of 6 occasions (once for the German test partition and twice for the Hungarian test partition). Further empirical work is required by the community to determine the best fusion technique for continuous affect prediction.

2.3.4.7 State-of-the-art affect prediction results summary and discussion

The publications referenced in Sections 2.3.4.1 to 2.3.4.6 are summarised in Table 2.1 and their continuous affect prediction test set results for arousal and valence are provided. It is noted that some of these results used additional modalities such

as electro-dermal activity or electrocardiogram features. However, the audio-video modalities likely contributed most to predictions as this is what annotators have access to and base their judgements on. The inclusion criteria for the table are that (1) The work has taken place in the last 5 years and (2) test set CCC results for arousal and valence prediction are provided as this, in theory[2], allows unbiased estimates of model performances. Cross-cultural works are not listed in this table as it is believed that a first step in developing relatively unexplored cues for continuous affect prediction is intra-corpus prediction. The listed publications are considered benchmarks against which the results presented in this work can be compared.

Input features for continuous affect prediction have traditionally been gathered based on expert-knowledge handcrafted efforts [7], [31], [113], from the paralinguistic/speech domain. However, there has been a trend toward using natural language processing [99] and deep learning techniques for this [34], [88] in recent times. This trend could ease what can be an extensive effort in time and resources for gathering handcrafted features, if successful. Unfortunately, these approaches do not conclusively outperform the handcrafted feature sets. For example, audio BoW [99] do not outperform functionals of handcrafted features [12] for audio only arousal prediction. Additionally, the best test set result in [99] for valence combined their BoW with handcrafted features. Further evidence of this, for the deep learned case, is given in [34]. While the authors outperformed unimodal handcrafted feature sets on the RECOLA [16] corpus and state their multimodal model greatly outperformed their unimodal ones, their multimodal system did not outperform [14] if unbiased estimates of model performance are considered from [34]. Moreover, no results in [34] outperform that of [12], who used handcrafted features, for arousal prediction. The multimodal results in [34] are superior to [14] if they consider models that perform best on the test set without matching (i.e. best) validation set performances. The results are biased because, contrary to the validation set results, the best performing models were selected based on test set performance. Both unbiased and biased estimates of model performance according to this author are provided from

[2]This is "in theory" due to potential for data leakage. For example, a legitimate test set pass contains gold standard annotation data not observable by the researcher in formulating their model according to Kaufman et al. [145] and can be considered an unbiased estimate of performance. However, if a model is considered based on test set *performance* with an accompanying learned model that is not the best performing model on observable inputs and targets (i.e. validation set), then data leakage has occurred. Here, a human has inadvertently used unseen data in the model selection process, which is a form of training, therefore providing a biased test set performance estimate. While this seems obvious, this can very easily happen to researchers who have otherwise good intentions. The result of this biased approach could lead to degraded model performance on future tests of the model were a set that more closely matches validation set variance is present, for example. Finally, test set results where data leakage has occurred are not completely invalid as they do show a score that could be achieved in modelling test set variance. However, they should not be taken as unbiased estimates of model performance. Where possible, biased estimates of continuous affect prediction model performance are highlighted in this work.

Table 2.1: Summary of State-of-the-art Continuous Affect Prediction Performance as Measured by Test Set CCC. Publications Are Listed From Highest Performing to Lowest in Terms of Average CCC (μ) Across Arousal and Valence.

Author(s)	ML Algorithm(s)	Data set	Arousal	Valence	μ	Notes
[14]	SVR, CNN & RNN, LSTM-RNN	RECOLA	.770	.687	.729	Kalman filter fusion, speech, face and physiology input
[13]	BLSTM-RNN	RECOLA	.747	.609	.678	Temporal feature to annotation offset, BLSTM-RNN fusion, speech, face and physiology input
[34]	CNN & LSTM-RNN	RECOLA	.715 (.789)	.620 (.732)	.668	End-to-end multimodal, (biased performance estimates), speech and image input
[12]	LSTM-RNN	RECOLA	.804	.528	.666	Individual annotator modelling, reaction-time context modelling, speech, face, head and physiology input
[88]	CNN	RECOLA	.603	.686	.645	Head pose and eye gaze guided attention, face, head and eye input
*[15]	SVR, BLSTM-RNN	RECOLA SEMAINE	.685 .346	.554 .026	.620 .186	Strength modelling, audio, video and face input
[99]	SVR	RECOLA	.753	.430	.592	Annotations delay, speech input only
*[128]	Multi-delay sinc CNN	RECOLA SEWA	.688 .412	.492 .379	.590 .396	Automatic annotator delay learning, speech input only
[144]	Gated recurrent unit RNN	RECOLA	.611	.527	.569	Implicit fusion, speech and image input
[119]	CNN & LSTM-RNN	RECOLA	.693	.352	.523	Annotation informativeness modelling, speech input only
[33]	CNN & BLSTM-RNN	RECOLA	.686	.261	.474	End-to-end learning, speech input only

*Used best corpus average for ranking

[34] in Table 2.1.

It can be seen from the literature review and Table 2.1 that head- and eye-based features are relatively unexplored for continuous affect prediction. Only Gunes and Pantic [72], Eyben et al. [125] and Ringeval et al. [12] used head-based features while Wu et al. [88] use cues from both of these modalities for affect prediction. This is despite comparable performance achieved by a head-based system compared to a speech-based system in [72] and the best performing arousal system from Table 2.1 using head-based features. While [88] used cues from both modalities, they do not focus on the eye gaze and head cues themselves but rather on how they can augment facial feature learning and prediction. It could also be argued that video cues provided for the AVEC challenges (e.g. [7], [8]) provide some information on head- and eye-based cues. However, the focus of these provided features is on facial expression and therefore, any head and eye information provided is in spite of, rather than due to, the feature engineering effort employed. Moreover, from [12], [72], [88], [125] it is difficult to discern the full extent that head- and eye-based cues can affect model performance. For example, comparisons against and combinations with alternative modalities are lacking and many more intra-modal features remain to be explored (e.g. mutual vs averted gaze, scanning vs fixated gaze, head-based frequency domain). It is clear that development of head- and eye-based cues for continuous affect prediction should be developed, based on results presented here and evidence to follow in the next section. Moreover, to optimise performance of these cues and to aid efforts in model interpretability, these cues should be developed initially based on expert/handcrafted feature methodologies.

When gathering input features for modelling algorithms, researchers should consider numerous temporal windows with a view to maximising the impact of extracted features [3], [99]. In this way temporal context may be captured for the affective expressions that annotators base their judgements on. Additionally, researchers should always consider, in some way, how features are temporally aligned with gold standard annotations to account for annotator reaction lag when providing ratings based on observed expression. This has been done explicitly, prior to modelling or using specifically designed algorithms, or implicitly, where the model used encodes the temporal saliency of features. Some researchers have used explicit temporal feature-to-target offsets by shifting forward features or shifting back annotations, even when using time-dynamic regression algorithms [13], [36], [127].

Feature selection approaches taken in the literature have largely used PCA [13], [121], [134] or correlation-based feature selection [13], [110], [131]. It appears that, for continuous affect prediction, consideration of nonlinear dependencies (or independence) might also be of benefit to the community, however, this has not yet been fully explored. For example, the mRMR algorithm [132] has performed well for con-

tinuous affect prediction [133] and in the cognate field of emotion recognition [146], and perhaps should be considered in future work on continuous affect prediction.

The community is not necessarily in agreement upon which affect learning and prediction algorithm is best to use, though LSTM-RNN type algorithms currently dominate the state-of-the-art. It has been questioned whether recurrence is required for affect prediction [129]. However, more research is needed to determine if CNN, for example, is a better algorithm than LSTM-RNN in terms of reducing model complexity and matching or exceeding performance. Schmitt et al. suggest that this can be the case [129]. However, this author believes that an even simpler feed-forward DNN architecture may suffice as the temporal window for feature extraction may capture enough context for continuous affect prediction. Of course, a DNN architecture would require explicit compensation for annotator reaction lag. However, this algorithm can produce a simpler overall model that would also be more suitable than BLSTM-RNN for real-time prediction. BLSTM-RNN's requirement of seeing full sequences prior to prediction make it problematic in this regard. CNN models are gaining in popularity for prediction [88], [128], [129], so perhaps further empirical evidence (for or against CNN) is forthcoming. Other algorithms employed include SVR, PLS, probabilistic output models (GMR, GPR, RVM, Kalman filter) and time-series approaches. Of note is the L1 regression approach used with these algorithms to train some of these models. This is something that this author believes should be investigated further for developing an interpretable affect prediction model. Finally, while the MTL approach for leveraging arousal and valence correlations during modelling has been used, it appears that it is not always effective [12] nor is the technique ubiquitous for modelling. Also, according to Table 2.1, arousal is generally predicted with higher fidelity, so there may be an opportunity to investigate a multiple-stage regression framework, inspired by teacher-forced learning [102, p. 377], for better valence prediction. In such a framework, arousal annotations (i.e. from the teacher) could be used to enhance training feature vectors and these features, or arousal validation set predictions, could be used in valence validation data.

In terms of fusion methods, there does not appear to be categorical evidence as to which of feature, decision or model fusion is best for affect prediction. For decision fusion, popular algorithms include LR and SVR while BLSTM-RNN networks are regularly used for model fusion. An interesting advantage to LR is that it promotes model interpretability, where learned weights correspond to final model contributions in a small search space. Furthermore, it has been suggested that simpler algorithms might outperform more complex algorithms in ensemble model learning [110]. It is perhaps therefore important to consider both of these algorithms for decision fusion while also considering feature fusion, or the most complex technique, model fusion,

for optimal modelling and appraisal of features.

This section has dealt with reviewing continuous affect prediction research related to this project. Important methodological steps have been identified for dealing with some challenges in learning and predicting affect (e.g. best input feature extraction method, annotator delay compensation). Moreover, areas for improvement within continuous affect prediction, such as considering head- and eye-based features, teacher-forced learning (multi-stage regression) and further pursuing non-linear feature selection have been identified.

2.4 Conclusion

This chapter has provided introduction and discussion across three core topics for this dissertation: affect/emotion measurement, head- and eye-based affective cues, and continuous affect prediction. Theoretically sound and important affect dimensions that make up core affect, arousal and valence, have been shown to be worthy of investigation in this work. Head- and eye-based features can be beneficial, based on the review of existing research, and these cues require exploration for continuous affect prediction. Addressing the use of these features for continuous affect prediction can allow a more holistic use of videos provided by subjects for this task. Feature sets are required to be developed for this task and further novel investigation of cross-modal feature interactions, including interactions with the popular speech modality, are required. Some ethical issues of the proposed modalities for continuous affect prediction have been discussed along with a licencing strategy and research application intended to disable mal-use of the research experiments and results to follow.

The review of continuous affect prediction literature informed methodological and algorithmic approaches taken for the experiments described in the chapters to follow. These include different temporal windows for feature extraction, techniques for annotator delay compensation, feature selection and the use of handcrafted feature engineering techniques, basing features on expert knowledge. The use of deep learning type algorithms for model generation is important to set this work fairly in context with that of the literature. Due to the complexity of the task of modelling affect, and based on works such as [12] and [139], different models (i.e. single task algorithms) were used for arousal and valence prediction in the experiments. Also, teacher-forced learning, implemented with multi-stage regression, was proposed as a novel technique for exploiting affect dimension correlations. The next chapter details the experimental data and methods used in this work.

Chapter 3

Affective Corpora and Experimental Approach

3.1 Introduction

This chapter details the affect data sets and the ML algorithm used in this work. Rationales for selections made are given along with exploration of data and explanation of algorithm operation where necessary. Finally, a high-level experimental architecture is provided that guides further detailed experimental designs to follow.

3.2 Data set selection

The data sets chosen for this work include RECOLA [16] and Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression (SEMAINE) [11] audio-video affect corpora. These corpora were selected due to (i) the natural affect displays present, (ii) differences in affect elicitation used, (iii) the high granularity of continuous arousal and valence annotations, and (iv) the strong experimental controls employed for data recording. For example, while the Vera am Mittag [147] corpus does contain arousal and valence ratings, only 5-point discrete Likert scale values in the range $[-1, 1]$ are provided. Moreover, the annotations in this corpus are provided at utterance and every third face image frame for audio and image corpus partitions respectively. This may be fine for turn-level affect recognition, but this is not the focus of this research, intended for pseudocontinuous affect prediction. The selected corpora contain continuous number ratings in the range $[-1.0, 1.0]$, with annotations provided in a discrete-time continuous manner based on audio-video, thus closer reflecting the true nature of affect which is both nuanced and continuously varying.

Only naturalistic affect data sets were chosen in this work because acted affect

data sets provide less of a challenge for learning algorithms, where the patterns to be observed can be exaggerated, and are often prototypical in nature [107]. In contrast to the exclusion of acted affective corpora, the SEWA corpus [53] was not selected as it may be overly challenging for developing previously underexplored affective features. SEWA [53] is a multilingual audio-video affective corpus and it has been recorded "in the wild", where different audio and video recording approaches have been used within the corpus. The AVEC 2014 [7] corpus was also considered, and preliminary affect prediction results were obtained on this corpus, but this set was ultimately excluded for two reasons. Less experimental recording control (lighting) is present in AVEC 2014 [7] and the corpus involves a human-computer interaction task, which may not reflect audiovisual communication as well as the selected corpora. The lack of lighting control on AVEC 2014 [7] made gathering direct gaze features, described in the next chapter, impossible for this work. The preliminary results gathered on AVEC 2014 [7] provided learning for the work carried out in this dissertation. These evaluations [148], [149] showed that eye features from video are worth investigating, and they can improve affect prediction when used with speech, compared to using speech alone. Another corpus that contains continuous arousal and valence ratings is HUMAINE [150], which was not selected for this work due to the lack of recent baseline results for comparison and the larger, less focused number of experimental settings in the corpus. Moreover, it should be noted that the selected corpus, RECOLA [16], is widely used for continuous affect prediction research, which facilitates the comparison of the results of this work against other recent research. The selected corpora are further explored in the sections to follow.

3.2.1 RECOLA corpus

The RECOLA [16] corpus was selected because it is widely used by the research community and allows for comparison of results with state-of-the-art approaches [12]–[14], [33], [34], [88], [99], [128]. The corpus contains audio, visual and physiological recordings of dyadic interactions between subjects collaborating on a task in French. The subjects within each dyad communicated remotely by way of computers and were required to reach consensus, while being recorded, on how to survive in a disaster scenario. This resulted in spontaneous affect displays during the social interaction. Subject meta-data such as subject age, sex and mother tongue are additionally provided. Recordings of 23 subjects available in the set were partitioned into training, validation and test sets with the aim of matching the distributions used in [12]. Specifically, the training set contained subjects {P16, P17, P19, P21, P23, P26, P30, P65}, the validation set included subjects {P25, P28, P34, P37, P41, P48, P56, P58}, and the test set included subjects {P39, P42, P43, P45, P46, P62,

P64}.

Continuous-valued annotations for arousal and valence in the range $[-1, 1]$ are provided frame-wise at 25 frames per second for each 5-minute recording in the data set. The video for the corpus is provided at the same rate as that of the annotation frequency. Individual annotation traces were provided for this corpus by 6 annotators (3 male and 3 female), where raters provided their perception of arousal or valence levels, separately, using a slider within a web-based interface. To obtain gold standard annotations from this corpus, in this work, the mean of all 6 raters was taken for each subject recording and affect.

3.2.2 SEMAINE corpus

The SEMAINE [11] corpus, of which a small subset of the data set was used, was selected to assess features developed in another similar, but different corpus, to provide additional validity for experimental findings. This data set contains audio and video, recorded using professional lighting and recording equipment, of subjects interacting in English with sensitive artificial listener (SAL) agents. The SALs included Solid SAL, a human playing a SAL agent, Semi-automatic SAL, where a human operator chose from a script to produce a virtual avatar audiovisual response, and Automatic SAL, a fully autonomous avatar. SAL characters were designed to give the impression that they were trying to make subjects feel angry, happy, gloomy or sensible. The subset of SEMAINE taken for the work presented in this dissertation comes from the Solid SAL scenario, in order to match the RECOLA [16] corpus (i.e. human-to-human discourse). Compared to free conversation, the only restriction was that subjects could not ask SAL any questions for the social interaction. The SAL reminded subjects of this if they did ask them anything. Subjects participated in these interactions as: operating SAL, interacting with SAL as a user of the system, or sometimes both (during separate session recordings) within the corpus. The subset taken for this work includes 6 males and 6 females and the data were partitioned into training, validation and test sets with the aim of balancing gender and SAL character interaction. Specifically, the subject recordings that comprised the training set were from sessions {S15, S21, S25, S34}, the validation set was comprised of sessions {S55, S59, S65}, and the test set {S72, S77, S84, S107}. The training and test sets were equally balanced in males, females and angry, happy, gloomy and sensible character interactions.

This corpus is provided with continuous-valued annotations for arousal, valence, expectation, intensity and power/dominance affect dimensions, provided at 50 frames per second. The video for the corpus matches the frame rate of the annotations. Annotations for arousal and valence in this corpus are provided in the range $[-1, 1]$

where the FEELtrace system [151] was used for gathering these ratings simultaneously. Each recording was rated by 2-8 annotators. The use of the FEELtrace system involved moving a cursor in an adjacent window to the video being listened to and viewed corresponding to their perception of the arousal and valence levels present in the video. The subset of SEMAINE [11] videos used in this research were all rated by 6 annotators. Gold standard annotations were taken as the mean of all of these annotator ratings for each subject recording and affect dimension.

3.2.3 Qualitative comparison of selected corpora

Differences in communication, affect elicitation and affective responses across the selected corpora are given in Table 3.1. Selecting corpora across different experimental protocols enables the assessment of the efficacy of proposed features for continuous affect prediction during different task-specific conditions and social scenarios. While the experimental corpora have beneficial differences, a potential confound is cultural differences between subjects. It is noted that it cannot be ruled out that one culture may have a preference for visual affect display as opposed to verbal, biasing the effectiveness of a particular modality.

Table 3.1: RECOLA and SEMAINE Audio-Video Corpora Language, Task, Communication and Affect Display Settings

Corpus	Language	Task	Communication	Affect
RECOLA [16]	French	Collaborative dyadic problem solving	Human to human	Natural, spontaneous
SEMAINE [11] Solid SAL	English	Responding to artificial character questions	Human to agent	Natural, elicited

3.2.4 Quantitative comparisons of selected corpora

For the following analyses, statistics for annotator agreement and arousal and valence annotations are provided. Visual distributions and correlations between arousal and valence are also given. Interesting similarities and differences between the corpora for these analyses are discussed.

3.2.4.1 Average annotator agreement

Average annotator agreement in CCC is provided as the mean CCC across annotator pairs in Table 3.2. These calculations were carried out for the entire data sets first, and then the on the individual partitions that contributed to the whole set analysis. If the average annotator CCC is taken as a baseline, the SEMAINE [11] corpus provides a high benchmark for automatic system prediction of valence in Table 3.2 compared to reviewed related works. It is suspected that the annotators

Table 3.2: Arousal and Valence Group-of-humans Average CCC (the mean CCC taken from all unique annotator-annotator rating pairs) for Each Data Set/Partition in the RECOLA and SEMAINE Corpora

Corpus	Partition	Arousal	Valence
RECOLA [16]	Whole set	.284	.364
	Training	.341	.383
	Validation	.293	.411
	Test	.217	.257
SEMAINE [11]	Whole set	.332	.611
	Training	.252	.501
	Validation	.384	.684
	Test	.398	.500

who provided the ratings for SEMAINE [11] are actually a group-of-experts as no training of annotators is mentioned in the corpus dissemination paper [11]. This may account for the high agreement in the valence traces. Both corpora appear to have comparable levels of average CCC across annotator pairs for arousal.

Of note for the SEMAINE [11] average annotator CCC calculations shown is that the S77 annotator group was not included in the whole set or test set analyses carried out. This was done because this session was annotated by a different group of annotators for which only one session sample/recording for that group was available. The calculated average annotator-annotator CCC scores for this session were 0.062 for arousal and 0.221 for valence. Incorporating these scores into the overall and test set group CCC scores would appear to lower the overall average but these scores were not incorporated as it was believed that this annotator group could not be represented by one sample/recording. This provides a unique challenge for the SEMAINE [11] test set in this work as model predictions are implicitly required to generalise to different annotator groups in addition to unseen input data.

It is interesting to note the higher average CCC across all corpora, in general, for the group-of-humans estimates for valence compared to arousal. For automatic systems, valence has traditionally been the harder dimension to predict [76] whereas for humans, based on the sample of data present, the reverse may be true. The values in Table 3.2 provide minimum practical performance measures and are used throughout this dissertation for comparison to validation and test set results achieved by automatic systems.

3.2.4.2 Arousal and valence statistics

Basic arousal and valence statistics for the corpora training sets are given in Table 3.3. It can be seen in both of these tables that there are a lot of unique ratings for the SEMAINE [11] corpus. This is perhaps due to the experimental protocol employed for SEMAINE [11], where the unique ratings may have resulted from the free-flowing style of conversation permitted with the various characters. Situations

such as this can present a challenge for automatic prediction systems where a lot of patterns have to be learned on the training data. This issue is profound in the SEMAINE [11] sample with 12,430 unique valence ratings present and over 13,000 unique arousal values.

Also, of note from these tables is the larger range of valence values in RECOLA [16] compared to SEMAINE [11]. The wider range of values can be beneficial, where systems can learn a wider variance in valence data patterns given enough training examples. This provides a challenging situation for valence model generalisation on SEMAINE [11] where less variance is available to learn from.

Table 3.3: RECOLA and SEMAINE Corpora Arousal (a) and Valence (b) Training Set Statistics Including Counts of Zero-rated Values (0s) and One-off (unique) Values

(a) Arousal

Corpus	min.	mean	median	max.	0s	unique
RECOLA [16]	-.595	.009	.053	.400	213	665
SEMAINE [11]	-.638	-.119	-.066	.361	0	13,103

(b) Valence

Corpus	min.	mean	median	max.	0s	unique
RECOLA [16]	-.248	.097	.088	.665	194	576
SEMAINE [11]	-.329	.112	.154	.481	0	12,430

3.2.4.3 Arousal and valence distributions and correlations

Arousal and valence distributions of each corpus are plotted in Figures 3.1 and 3.2. Respective Pearson's r values for the corpora arousal and valence are 0.616 and 0.755 for RECOLA [16] and SEMAINE [11]. The correlations quantitatively illustrate that there are linear relationships across these core affect dimensions that could be leveraged for modelling. In the case of SEMAINE [11] there is a clear group of very low valence values that might be considered outliers, which is biasing the correlation to a low value. It can also be observed in the figures that the corpora seem positively biased in the valence polarity distribution.

3.3 Machine learning algorithm

Extensive use has been made of SVR and LSTM-RNN-type networks for continuous affect prediction [12], [14], [34], [36], [131]. The current popularity of LSTM-RNN-type networks provides for temporal alignment of gold standard target values with input features. However, this comes at the expense of a complex model. Another disadvantage of BLSTM-RNN specifically, is that the entire sequence must be seen by the model prior to prediction so it can consider both future input values in addition to past and current input. Moreover, while it has been argued that LSTM-RNNs can capture the temporally salient features in affect [135], the temporal window for

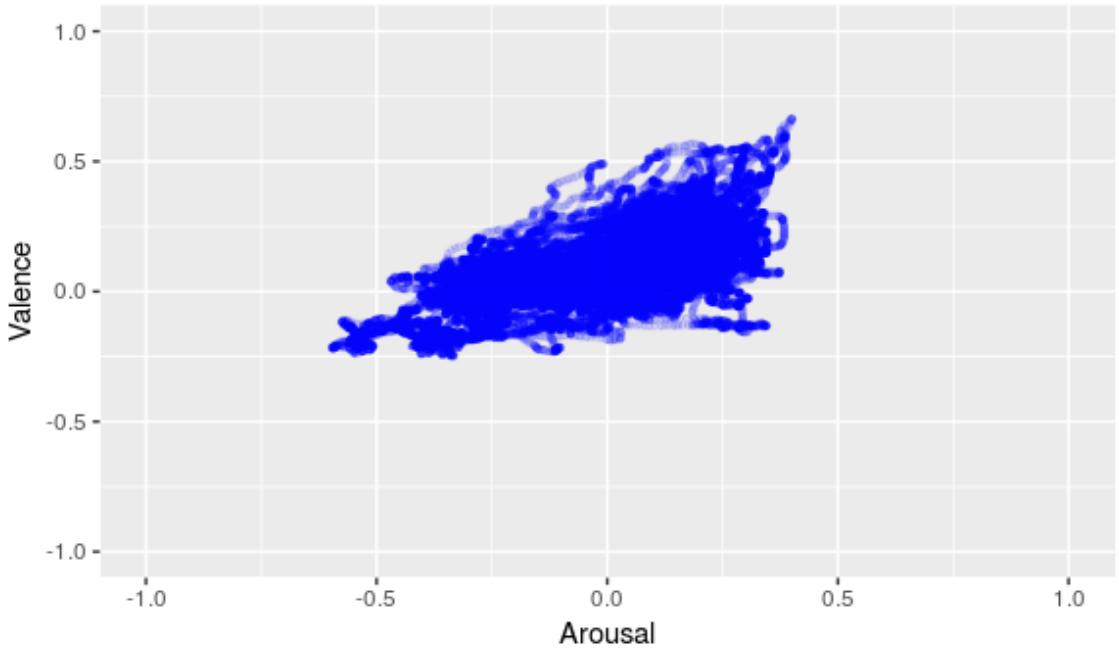


Figure 3.1: RECOLA arousal-valence scatter plot. $r = .616$.

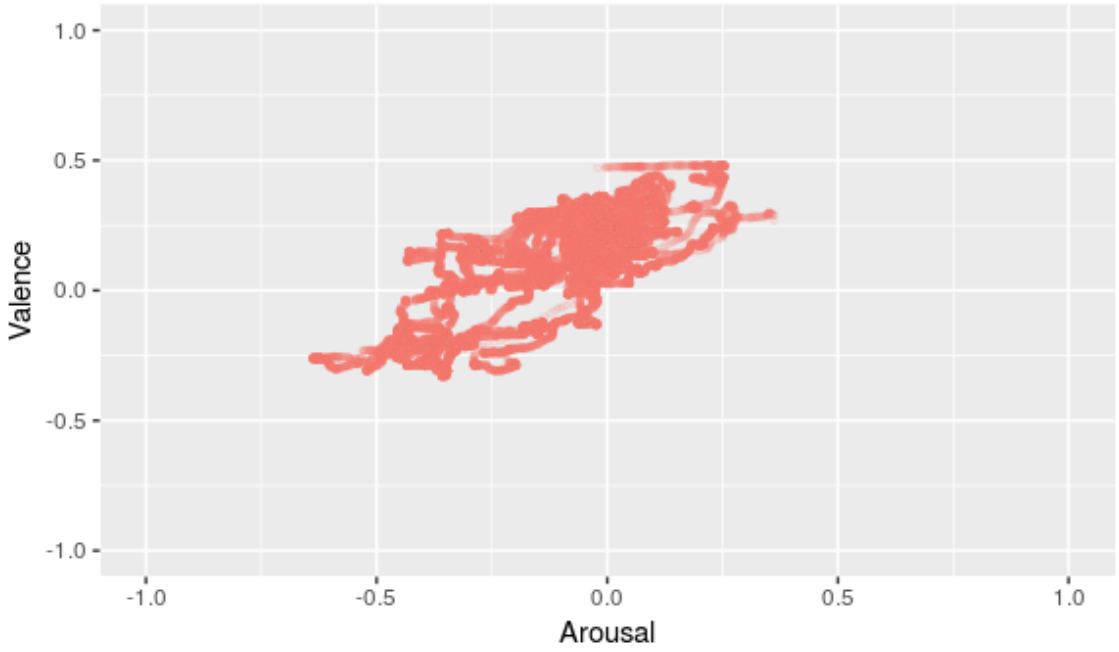


Figure 3.2: SEMAINE arousal-valence scatter plot. $r = .755$.

feature extraction may also capture this context if large enough. With this being said, LSTM-RNNs, and DNNs generally, facilitate multi-level feature interactions at varying levels of depth, which may be important in multimodal tasks such as recognising affective expression. Therefore, as a bridge between SVR and LSTM-RNN, DNNs were used in this work, along with explicit gold standard annotator lag compensation. Due to the lack of wide use of feed-forward type DNNs for continuous affect prediction, this work provides a contribution in affect modelling for the community. Specific details on the DNN architecture and training method used are provided in the sections to follow.

3.3.1 Architecture

The DNN architecture used is the same as Ringeval et al. [12] for their feed-forward neural network evaluations. Specifically, a DNN with two hidden layers of 160 and 120 hidden nodes were trained and evaluated with single-task learning (i.e. arousal or valence learning and prediction) using the Cuda recurrent neural network toolkit (CURRENNT) [152]. A tanh function was used for hidden layer activations where the forward equations are:

$$\mathbf{h}_1 = \tanh(\mathbf{W}_1^\top \mathbf{x} + \mathbf{b}_1); \tag{3.1}$$

$$\mathbf{h}_2 = \tanh(\mathbf{W}_2^\top \mathbf{h}_1 + \mathbf{b}_2), \tag{3.2}$$

where $\mathbf{x}$ is an input feature vector, $\mathbf{h}$ are hidden layer activations, and $\mathbf{W}$ and $\mathbf{b}$ are weight matrices and bias vectors respectively containing parameters to be learned. Finally, $\hat{y}$ is provided in the output layer by a linear activation, specifically, a summation of $\mathbf{h}_2$ and a further bias term

$$\hat{y} = b + \sum_i h_i, \tag{3.3}$$

where h_i are elements of $\mathbf{h}_2$.

3.3.2 Network training and model selection

The DNN training methodology employed follows that of Ringeval et al. [12] with the only difference being that the early stopping parameter was set to 10 instead of 20. The justification for this alteration was based on faster network training time and early experimentation showing no performance increase on the validation set for the larger early stopping time parameter. Prior to network training, weights were initialised to Gaussian distributed random numbers with a mean of 0 and standard

deviation of 0.1. Biases were initialised to 1.0. Inputs and gold standard targets were standardised to zero mean and unit variance prior to training, using means and standard deviations computed on the whole training partition. In addition, Gaussian noise with a standard deviation of 0.1 was added to all input features prior to training as an effort to prevent overfitting. Training took place for a maximum of 100 epochs using a sum of squared errors objective in mini-batches of size 10. Stochastic gradient descent with momentum of 0.9 was used for optimisation:

$$\hat{g} = \frac{1}{m} \sum_{i} \partial_{\theta} L_i, \tag{3.4}$$

$$v = \beta v - \alpha \hat{g}, \tag{3.5}$$

$$\theta = \theta + v, \tag{3.6}$$

where $\hat{g}$ is the gradient estimate for a batch of m training examples, L_i is a training example loss, v is the velocity, θ is a parameter to be learned, α is the learning rate and $\beta = 0.9$ is the momentum hyperparameter. This hyperparameter controls how much an exponentially-weighted moving average of gradients, v, affects the current gradient. Larger v and smaller α results in higher weighting of previous gradients compared to the current gradient in the update. Momentum speeds up gradient descent and provides more reliable steps (less variance) toward optimum parameter values. Training stopped if no improvement of the performance as measured by sum of squared error was observed on the validation set for more than 10 epochs for further regularisation. The network learning rate was held at the default value in CURRENNT [152], 10^{-5}, for all the experiments. Model selection was always performed based on models that provided the best validation set performance from the experiments.

3.4 Experimental architecture

The experimental architecture is intended to provide thorough and fair evaluation of the proposed features at each experimental stage. A simple visualisation of this process is shown where features are extracted and models trained in Figure 1.1c. The overall architecture in high-level terms for the experiments in this dissertation is as follows:

- Initial feature set proposal(s) for head- eye-based cues based on theory or prior empirical results.

- Unimodal modelling using ML algorithms.
- Multimodal modelling using ML algorithms and modality fusion.
- Performance evaluation of trained model performances using CCC.
- Test set passes for the final developed models and retrospective inspection and discussion of head and eye features used for final models.

The experimental architecture was chosen based on the previously reviewed related work, where models were trained to learn dimensional and continuous affect. The chosen performance metric, CCC, as evident from the related research, is now the de facto performance metric for continuous affect prediction. Finally, for the retrospective, measures of feature-with-target (arousal and valence) dependence are provided for selected features that were used as input to the final models. Details on the specific experiment steps taken at each phase of this work are given in the chapters to follow. Additional information on software used for the experiments is provided in Appendix A.

3.5 Conclusion

This section has detailed the data and some methodological steps taken in this work. Data used has been presented and explored with some similarities and differences shown across selected data. This provides technical challenges and opportunities, such as modelling rarely observed or unique annotation values, or effectively leveraging correlations between arousal and valence. The high-level experimental architecture was presented and serves as a guide for more detailed experiments in the chapters to follow. The next chapter details the exploration of LLDs and the proposal of feature sets, from head and eye modalities, for continuous affect prediction.

Chapter 4

Feature Set Proposals and Unimodal Evaluations

4.1 Introduction

Eye gaze has been shown to be important for emotion signalling and perception in the literature [22], [23], [59], [63] while eye pupils are responsive during emotional arousal [27], [28] and events of various valences such as monetary incentive or penalty [29]. Head pose is also important for emotion recognition both in the presence and absence of speech [24], [26], [59], [62], [65]. For head-based affect prediction, assessing frequency domain measures is important as humans have been shown to be sensitive to these cues while detecting emotion in utterances [65]. Furthermore, some eye cues share a relationship with head pose [59], [67], [153] as certain gazes require particular head poses. Feature sets from video that use these modalities for continuous affect prediction are currently lacking, despite their ease-of-use and potential usefulness. Motivated by these facts, this chapter presents five novel feature sets, three for eye-based features and two for head-based features. The feature set proposals were followed by unimodal experiments intended to assess their usefulness. Specifically, the identified research question for this chapter is

How well do head- and eye-based features perform compared with speech and facial features for unimodal continuous affect prediction?

Deep learning experiments conducted on the RECOLA [16] and SEMAINE [11] corpora showed the proposed head-based features to be the best-performing visual arousal feature set from those evaluated, with speech performing best overall. The head and eye features performed better than speech and face for valence prediction on SEMAINE [11]. This was not replicated on RECOLA [16], where face feature

performed best for valence prediction. These results confirm that speech performs best for unimodal arousal prediction while visual features are best for valence prediction from audiovisual data, results commonly known in the literature. The results also indicate that head-based features can be of benefit for arousal prediction from visual features.

4.2 Feature sets

Because of the potential that head- and eye-based features from video have for affect prediction, this section presents details of processes carried out to calculate LLDs from this data source and modalities. Explorations of these LLDs are also carried out before a proposal for affective feature sets from head- and eye-based cues are given.

4.2.1 Feature set LLDs and exploratory analyses

4.2.1.1 Eye LLDs

The eye-based LLDs were calculated from, or based wholly on, world or camera coordinates obtained from video using OpenFace [92] (version 2.0.6). They were gathered frame-wise from each subject's video recording. Figure 4.1 depicts some raw data points gathered. Eye blink/closure, blink intensity and x and y gaze angle (radians) LLDs were taken directly from OpenFace while Δs of x, y gaze angles, eye fixation and eye gaze approach required further calculation. OpenFace AU detection, namely, AU45_c and AU45_r, was used to determine eye blink/closure and eye blink intensity respectively. For these LLDs, _c is the binary presence or absence of an AU and _r is the real-valued intensity of an AU (from 0.0 to 5.0). OpenFace [92] provides gaze angle radian estimations, where x or y are the average angles for each coordinate across both eyes. Therefore, eye fixation was determined by examining the absolute value of the frame-wise change of the x and y radian measurements. When this measurement did not exceed the allowed fixation drift, small eye movements within a fixation event, which can be as much as $3.2°/second$ according to Lappi [67] fixation was determined as true (= 1) and otherwise, false (= 0). The degree figure from Lappi [67] was transformed to frame-wise radians ($0.05585 rad$/frames per second) for the experiments presented here. In the case of RECOLA [16], which was recorded at 25 frames per second, for example, this equated to 0.002 radians of allowed movement for each axis during fixation, when rounded to 3 digits. The gaze distance was calculated as the average distance on the z axis between each of the 2 outer-eye corner landmarks highlighted in green in

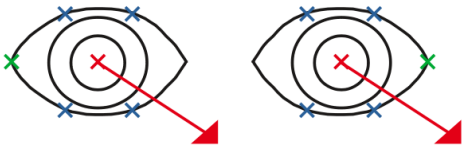


Figure 4.1: Eye gaze data points provided by OpenFace [91] for eye gaze vectors (red) and gaze distance (green) estimation.

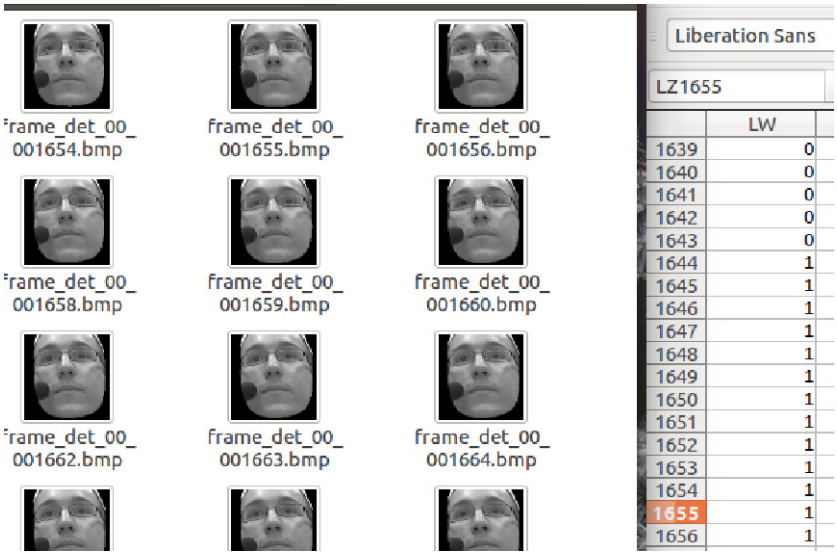


Figure 4.2: Direct (1) and averted (0) gaze annotation based on frame images.

Figure 4.1. From the gaze distance measurement, eye gaze approach was determined based on the frame-wise change in z, Δz, being positive.

An additional eye gaze LLD, direct gaze, where a subject was looking at an interlocutor, was also extracted. Healthy humans are adept at paying attention to the eye area of other humans [68] with some amygdala functions dedicated to eye area attention [30], [71]. For the direct gaze LLD gathered for the experiments, binary ratings for direct gaze (= 1) and averted gaze (= 0) were manually assigned by this author for every frame in the RECOLA [16] corpus. These binary annotations were assigned in the same way to the SEMAINE [11] corpus subset. This process was carried out based on frame images from OpenFace [92] output as shown in Figure 4.2.

Pupil diameter estimations were obtained from OpenFace [92] eye landmark coordinate differences. The eye landmarks used for pupil diameter estimation are shown in Figure 4.3. From the eye landmarks shown in the figure, the pupil diameter estimation was taken as $\max(x_2 - x_1, y_2 - y_1)$ for each frame. Also, only one pupil was used as the estimator for both eyes as both pupils should be the same in healthy subjects [79]. The pupil diameter estimations generated a number of pupil LLDs, namely, numerical pupil diameter and Δ pupil diameter along with binary dilation and constriction event data. The latter LLDs are based on pupil diameter

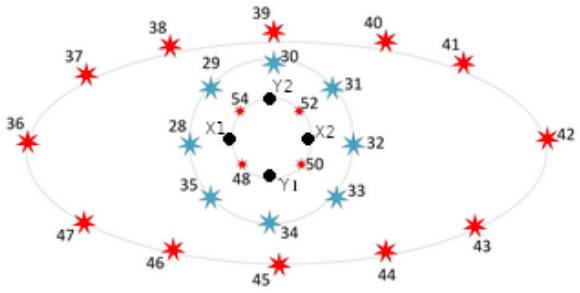


Figure 4.3: Eye landmark data points provided by OpenFace [92]. Data points used for pupil diameter estimation are highlighted in black.

measurements in each frame becoming larger or smaller.

For the exploratory analysis, descriptive statistics were compiled for the frame-wise LLDs on the training partition recordings in order to aid general data understanding and quality. Following this, the LLD features' relationships with each other, and with both arousal and valence, were explored using correlation and MI analysis. Cover and Thomas [154] describe MI, $I(X;Y)$, as the entropy between the joint distribution $p(x,y)$ and the product distribution $p(x)p(y)$ of two random variables, where $p(x)$ and $p(y)$ are marginal probability mass functions, as described by Equation (4.1). The mutual information between vectors $\mathbf{x}$ and $\mathbf{y}$ is the average information that is known about $\mathbf{y}$ given $\mathbf{x}$, expressed in nats in this work.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{4.1}$$

For the mutual information estimation, features and target data were first discretised into $N^{\frac{1}{3}}$ bins to facilitate $p(x,y)$ estimation. MI was then estimated between the features and arousal, and the features and valence, using the empirical distribution method. In this way, both linear and nonlinear feature-with-feature and feature-with-target relationships were explored, which is required for full exploration of dependencies, as, for example, $r = 0.0$ does not imply independence. As an upper-limit reference point, maximum MI estimated in this work, the information of a variable with itself, was 3.638 nats on RECOLA [16] and 3.584 nats on SEMAINE [11]. Temporal averaging of the LLDs, applied with 4-, 6- and 8-second windows, moved forward at a rate of 1 frame per interval, were used for this analysis based on [3]. This resulted in, for example, a 99% overlap for contiguous feature chunks when using a 4-second temporal window on data recorded at 25 frames per second. An example diagram of this frame chunking into the temporal windows is provided in Figure 4.4.

The results from the descriptive statistics analysis can be seen in Tables 4.1 and 4.2. It can be seen in Table 4.1 that there are not too many unique values for any of

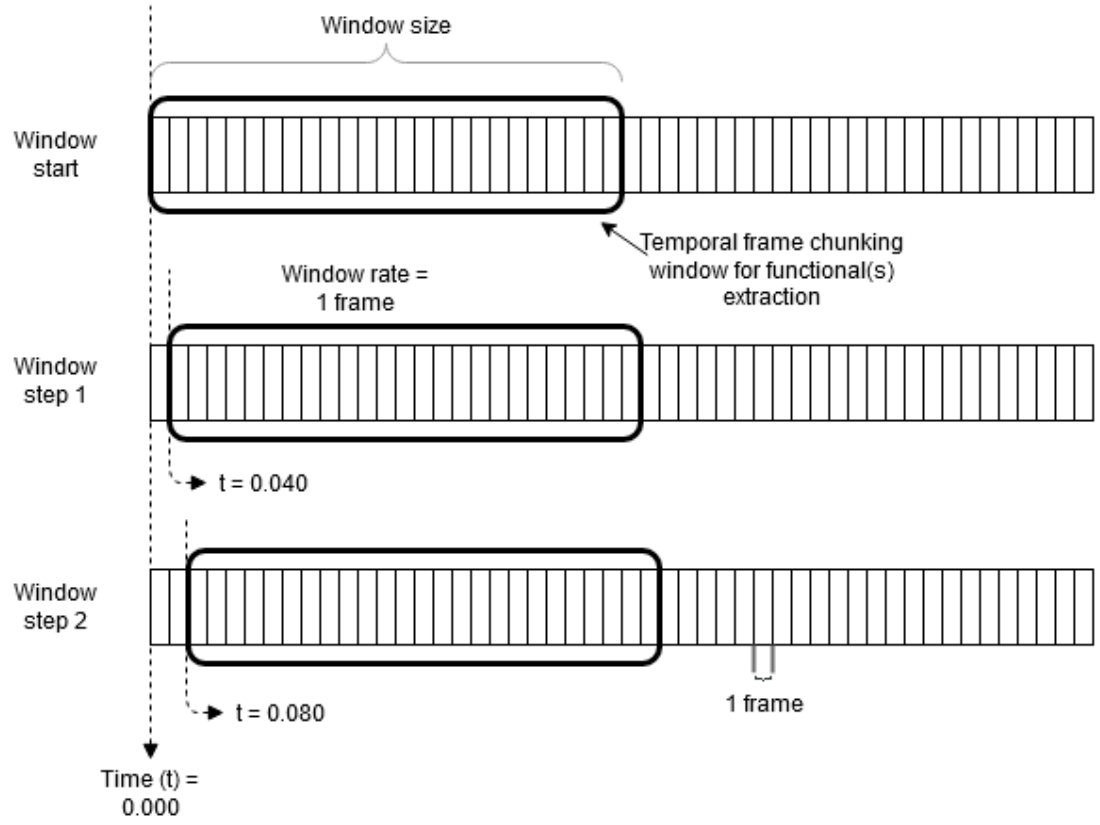


Figure 4.4: Chunking of frames into temporal windows for average (or other functionals, if required) extraction. The window size shown in bold this diagram is 1-second in size, based on a sampling rate of 25 frames per second, to facilitate illustration. At each step shown in the diagram, it can be observed that the window advances in time by a rate of 1 frame.

the numeric variables on either corpus. This table also shows that, due to the one-sided nature of eye blink intensity, gathering functionals such as minimum values are unwise for this LLD. The pupil diameter LLD has the least unique values of the eye descriptors across both corpora. This seems reasonable as the subjects should most likely all take on ranges of similar values for their pupil sizes [79]. From Table 4.1 (a) it can also be observed that some questionable values were obtained where pupil diameter estimation failed on RECOLA [16]. For example, a pupil diameter value of 0.1 mm is not a normal value [79]. Based on this, the proportion of questionable values for pupil diameter on RECOLA [16] was calculated, defined as the number of values in this set less than 2 mm, based on [79] which defines the possible range of values as 2 to 8 mm. In total, these questionable values only made up 0.33% of the total pupil diameter values in the RECOLA [16] corpus training set. This was deemed as acceptable noise.

The binary LLD descriptives in Table 4.2 indicate that eye blink/closure is relatively sparse, as would be expected based on a healthy subject's blinking rates [69].

Table 4.1: Numeric Pupil LLD Statistics Including Counts of Zero-rated Values (0s) and One-off (unique) Values Calculated on the (a) RECOLA and (b) SEMAINE Training Partitions

(a) RECOLA

LLD	min.	mean	median	max.	0s	unique
x gaze angle	-.776	-.035	-.009	.694	6,419	928
y gaze angle	-.926	.270	.331	.973	6,309	1,249
Δx gaze angle	-.702	.000	.000	.943	7,841	1,428
Δy gaze angle	-.739	.000	.000	.926	7,975	2,060
eye blink intensity	0.000	0.260	0.000	5.000	34,105	456
pupil diameter (mm)	0.100	5.480	5.500	8.200	0	456
Δ pupil diameter (mm)	-6.200	0.000	0.00	6.500	9,090	1,226

(b) SEMAINE

LLD	min.	mean	median	max.	0s	unique
x gaze angle	-0.351	0.044	0.040	0.403	341	690
y gaze angle	-0.288	0.091	0.093	0.642	179	764
Δx gaze angle	-0.237	0.000	0.000	0.340	2,765	850
Δy gaze angle	-0.358	0.000	0.000	0.293	2,634	293
eye blink intensity	0.000	0.300	0.000	3.900	25,596	381
pupil diameter (mm)	3.800	5.410	5.400	7.5	0	195
Δ pupil diameter (mm)	-2.400	0.000	0.000	2.600	9,900	412

Eye blink, which is eye closure of 0.1 to 0.4 seconds[1], compared to eye closure (eye closed $>$ 0.4 seconds) was not discriminated in this work due to the potential aversive signalling component of eye closure events [32]. Eyes fixated was even sparser than eye blink/closure, with its presence occurring in 11.4% of frames on RECOLA [16] and only 0.9% on SEMAINE [11]. This is thought to be reasonable for RECOLA [16] given the dyadic problem interaction task in question where multiple gaze shifts and searches were necessary, however, this proportion seems very small for SEMAINE [11]. Table 4.2 shows that direct gaze is less frequent than averted gaze in the RECOLA [16] corpus sample while the reverse is true for SEMAINE [11]. The differences across the corpora in terms of these gazing patterns is interesting as they may provide clues on the social context for video recordings.

The results of the feature-with-feature and feature-with-target dependency analysis showed that larger temporal windows performed best for feature Pearson's r and MI (in nats) with both arousal and valence on both corpora. On RECOLA [16], the largest correlations with arousal and valence were achieved with x gaze angle (arousal $r = 0.261$, valence $r = 0.242$) for the 8-second temporal window condition. The largest MI in this corpus for both arousal and valence was also observed for this descriptor with the same temporal window, MI values of 0.339 and 0.346 for arousal and valence respectively. On SEMAINE [11], again, the largest relationships were all found for the 8-second temporal window condition. Specifically, eye blink/closure had the largest correlation with arousal, $r = 0.181$, and y gaze angle shared the largest linear relationship with valence, $r = 0.495$. The y gaze angle LLD arousal had the strongest nonlinear relationships with both arousal and valence on

[1] bionumbers.hms.harvard.edu/bionumber.aspx?id=100706&ver=4&trm=blink+frequency&org=

Table 4.2: Binary Eye-based LLD Statistics (where absence and presence are indicated by 0 and 1 respectively) Calculated on the (a) RECOLA and (b) SEMAINE Training Partitions

(a) RECOLA

LLD	0s	1s	%1s
eye blink/closure	43,124	16,578	28.1%
eyes fixated	52,706	6,996	11.7%
eye gaze approach	30,243	29,459	49.3%
direct gaze	46,440	13,262	22.2%
pupil dilation	34,502	25,200	42.2%
pupil constriction	34,290	25,412	42.6%

(a) SEMAINE

LLD	0s	1s	%1s
eye blink/closure	33,187	14,506	30.4%
eyes fixated	47,274	419	0.9%
eye gaze approach	23,771	23,922	50.2%
direct gaze	15,237	32,456	68.5%
pupil dilation	28,832	18,861	39.5%
pupil constriction	28,761	18,932	39.7%

SEMAINE [11], an arousal MI of 0.861 and valence MI of 0.906. These results provide early indication that functionals of gazing angle LLDs may be relevant for predicting affect from eye gaze. Further, they show differences across the corpora, where the left-to-right gazing angle appears to be a more important affective signal in RECOLA [16] while the up-down gazing angle appears to be more important in SEMAINE [11].

The strongest positive feature-with-feature linear relationship found on RECOLA [16] included eye gaze approach and pupil constriction (r = 0.803), and the largest negative relationship was y gaze angle and direct gaze (r = -0.640), for 8- and 4-second temporal windows respectively. The y gaze angle and direct gaze LLDs shared the largest feature-with-feature nonlinear relationship under the 8-second temporal window condition, a MI of 0.578. On SEMAINE [11], the largest feature-with-feature linear relationships included pupil constriction and pupil dilation (r = 0.819), from the positive associations, and direct gaze and eye blink/closure (r = -0.678) from the negative associations. Eye blink intensity and eye blink/closure shared the largest nonlinear relationship, a MI of 0.902. It could be argued that there is feature redundancy between eye gaze approach and pupil constriction on RECOLA [16], and pupil constriction and pupil dilation on SEMAINE [11]. However, these LLDs were retained in the set for feature extraction and machine learning processes.

Heat map results of these analyses can be observed for the 8-second temporal window in Figures 4.5 and 4.6 where some of the aforementioned relationships can be visually observed. According to Figure 4.5, the eye features are more linearly relevant for arousal and valence on the RECOLA [16] corpus, but there appears to be stronger feature redundancies on that set. The stronger nonlinear relationships

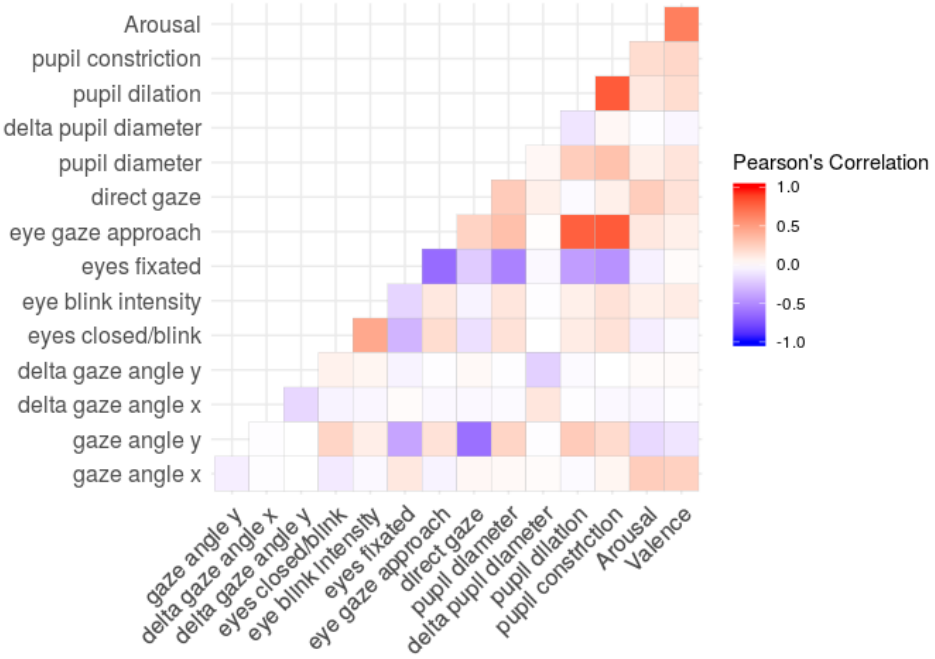
for the eye features with arousal and valence, however, appear on the SEMAINE [11] set (see Figure 4.6), again with stronger feature-with-feature relationships, or, redundancies. Overall, this analysis reveals that the larger temporal window, despite providing some information loss (smoothing), appears to be the best for feature representation and prediction. This is because the larger window most often produces the largest feature-with-target r and MI. Additionally, in some cases, the information reduction caused less feature-with-feature correlation, which can benefit for learning algorithms as more unique information can be leveraged for pattern learning.

4.2.1.2 Head pose LLDs

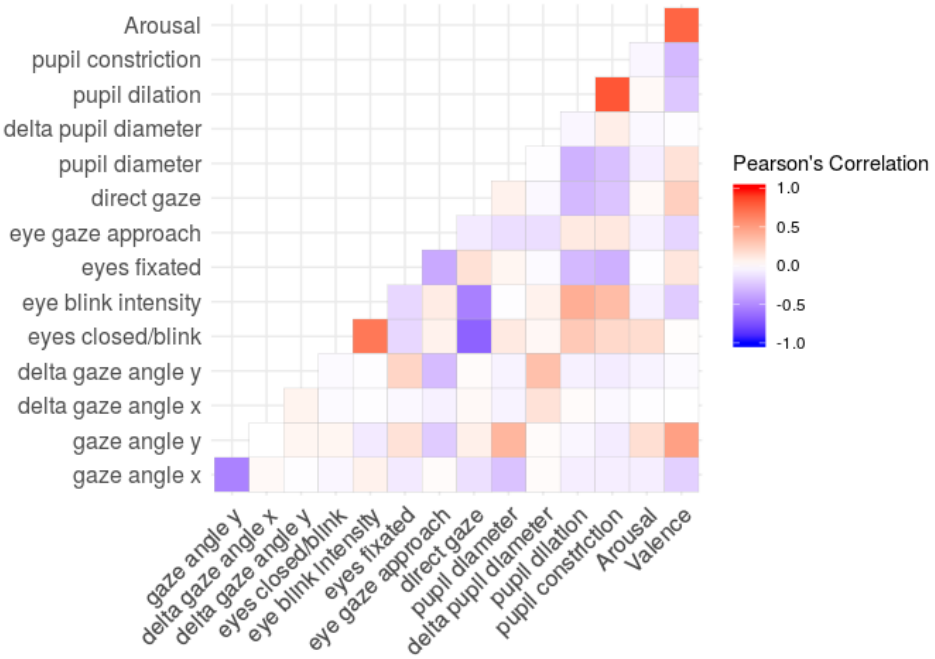
The head pose LLDs selected for this work include rotational and location measurements of head pose that are estimated from video using OpenFace [92]. Frame-wise displacements (Δ) were also calculated and included in the LLD list. Velocities were considered as additional LLDs but due to the additional computational cost and high correlation with displacement, it was decided against including head movement velocities at this time. The feature set LLDs are x (pitch), y (yaw) and z (roll) of head rotation in world coordinate radians with camera origin, and x, y and z of head location in camera coordinate millimetres. The descriptives for these LLDs can be seen in Table 4.3.

Table 4.3 indicates that performing feature scaling to put head pose features within similar ranges is important as vastly different scales can be observed when comparing rotation versus location LLDs. There are low amounts of unique values in the table which suggests that there is adequate but not excessive variation in the head pose data. More variation in the data is present for the raw x, y, z measurements as opposed to the Δs, evident by the much lower number of zeros present. This is expected as differencing can often produce 0 values while additionally introducing a more stable mean. An advantage of the Δ features is that they highlight high-frequency components of the signal, which could be advantageous if the true affective signal is of a high-frequency nature. In comparing sections (a) and (b) of Table 4.3, it can be observed that there are a wider range of head movements on RECOLA [16] compared with that of SEMAINE [11]. The differing interaction tasks across the corpora may have required different head poses from subjects. It will be of interest to see if these differences have an effect on prediction system performance.

The feature-with-feature and feature-with-target dependency analyses of the head LLDs were again carried out using the same temporal windowing methods as that for the eye-based exploration. It was again found that larger temporal windows were always better for feature-with-target r and MI. The strongest correlations

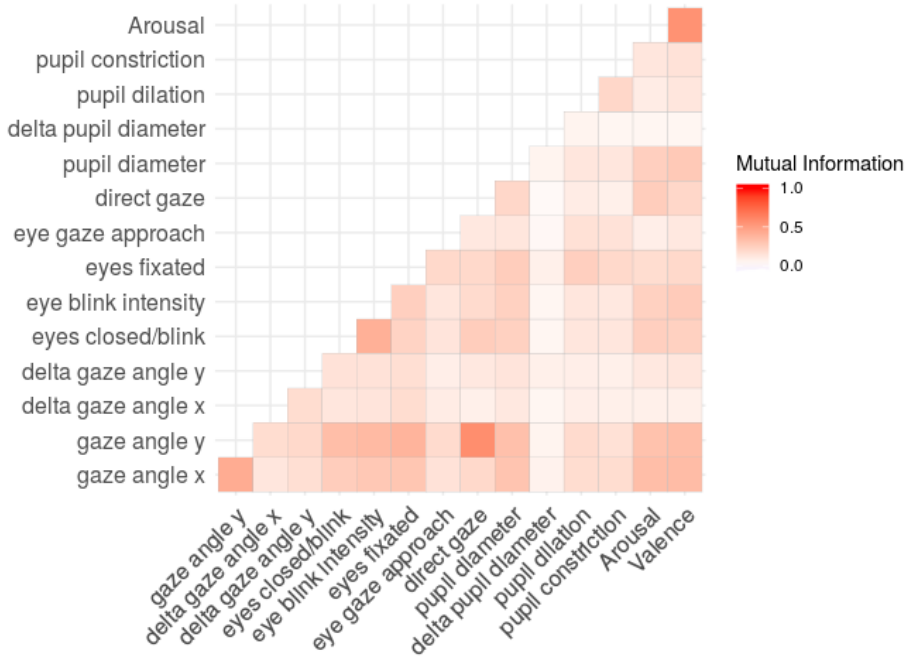


(a)

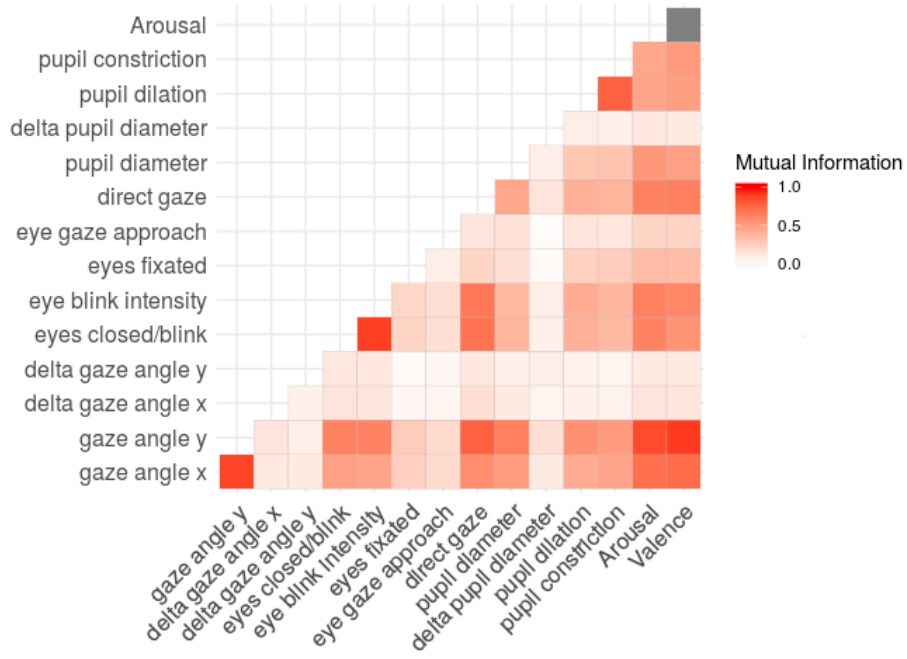


(b)

Figure 4.5: Eye correlation heatmaps for 8-second moving average of LLDs on (a) RECOLA and (b) SEMAINE.



(a)



(b)

Figure 4.6: Eye MI heatmaps for 8-second moving average of LLDs on (a) RECOLA and (b) SEMAINE. Gray-coloured tiles indicate MI > 1.

Table 4.3: Head Pose LLD Statistics Including Counts of Zero-rated Values (0s) and One-off (unique) Values Calculated on the (a) RECOLA and (b) SEMAINE Training Partitions

(a) RECOLA

LLD	min.	mean	median	max.	0s	unique
x location	-553.100	4.070	-5.100	259.900	60	2,614
y location	-2,218.700	40.280	38.600	1,050.600	35	2,997
z location	-2,498.800	407.580	421.300	1,552.000	0	4,499
x rotation	-2.750	0.127	0.131	2.722	61	2,143
y rotation	-1.426	0.020	0.027	1.495	141	1,956
z rotation	-2.938	0.108	0.089	3.056	119	1,840
Δx location	-583.700	0.000	0.000	516.600	8,585	1,171
Δy location	-2,160.200	0.010	0.000	1,964.900	8,470	1,533
Δz location	-2,049.600	0.000	0.000	2,333.600	3,446	1,436
Δx rotation	-2.773	0.000	0.000	2.750	4,973	1,935
Δy rotation	-2.164	0.000	0.000	1.500	7,036	1,428
Δz rotation	-2.848	0.000	0.000	5.994	9,186	1,380

(b) SEMAINE

LLD	min.	mean	median	max.	0s	unique
x location	-156.000	-50.460	-44.200	38.900	1	1,369
y location	-22.400	68.540	65.200	132.100	0	1,036
z location	156.600	432.38	448.300	547.000	0	2,351
x rotation	-0.468	-0.048	-0.041	0.627	107	848
y rotation	-0.452	-0.026	-0.016	0.458	381	777
z rotation	-0.345	0.022	0.003	0.426	209	663
Δx location	-25.600	0.000	0.000	20.100	7,636	439
Δy location	-12.900	0.000	0.000	10.200	6,898	322
Δz location	-65.000	0.000	0.000	146.200	2,589	464
Δx rotation	-0.492	0.000	0.000	0.193	4,736	618
Δy rotation	-0.257	0.000	0.000	0.218	7,332	498
Δz rotation	-0.092	0.000	0.000	0.086	10,513	478

with both arousal and valence on RECOLA [16] resulted from the x head location LLD feature (arousal r = -0.244, valence r = -0.142) for the 8-second window. The largest MI on this corpus with both arousal and valence resulted from z head location, values of 0.505 and 0.526 respectively, again with the 8-second window. For SEMAINE [11], all the strongest relationships were observed with the 8-second window condition. Specifically, the strongest correlation with arousal resulted from y rotation (r = 0.177), while the strongest correlation with valence resulted from y location (r = -0.449). The largest nonlinear relationships with both arousal and valence on SEMAINE [11] was from z location (arousal MI = 1.098, valence MI = 1.109).

Some strong feature-with-feature relationships among the LLDs on RECOLA [16] included Δ y location and Δ z location (r = 0.542), y location and z location (r = 0.474, MI 0.802), and x location and z location (MI = 0.831). Strong feature-with-feature relationships observed on SEMAINE [11] were z location and x location (r = -0.761), x location and y location (r = -0.662), z location and x location (MI = 1.271), and z location and y location (MI = 1.284). As a general observation, a higher quantity of noticeable relationships were observed for feature-with-feature head LLDs compared to the eye-based LLDs. No LLDs were dropped from the feature set prior to further feature extraction. The results from the best

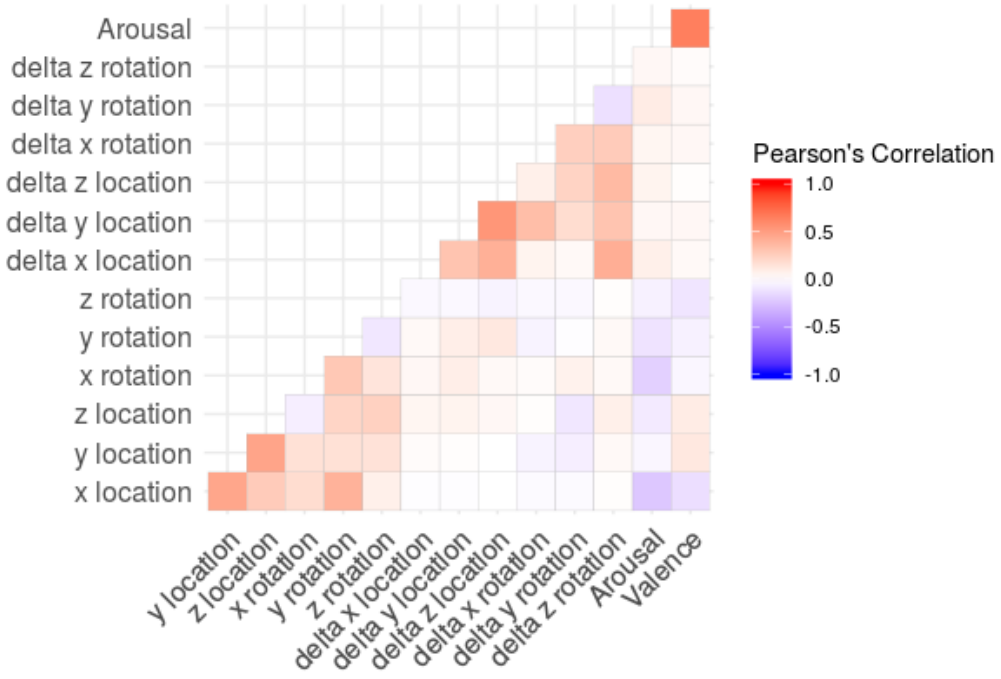
performing 8-second temporal window, in terms of feature-with-target r and MI are provided in Figures 4.7 and 4.8. In the figures, it can be observed that there are relatively similar feature-with-feature and feature-with-target relationship strengths across the corpora for the head-based LLDs. The polarity, however, appears reversed for the linear relationships across similar variables on the different corpora depicted in Figure 4.7.

4.2.2 Mid-level features

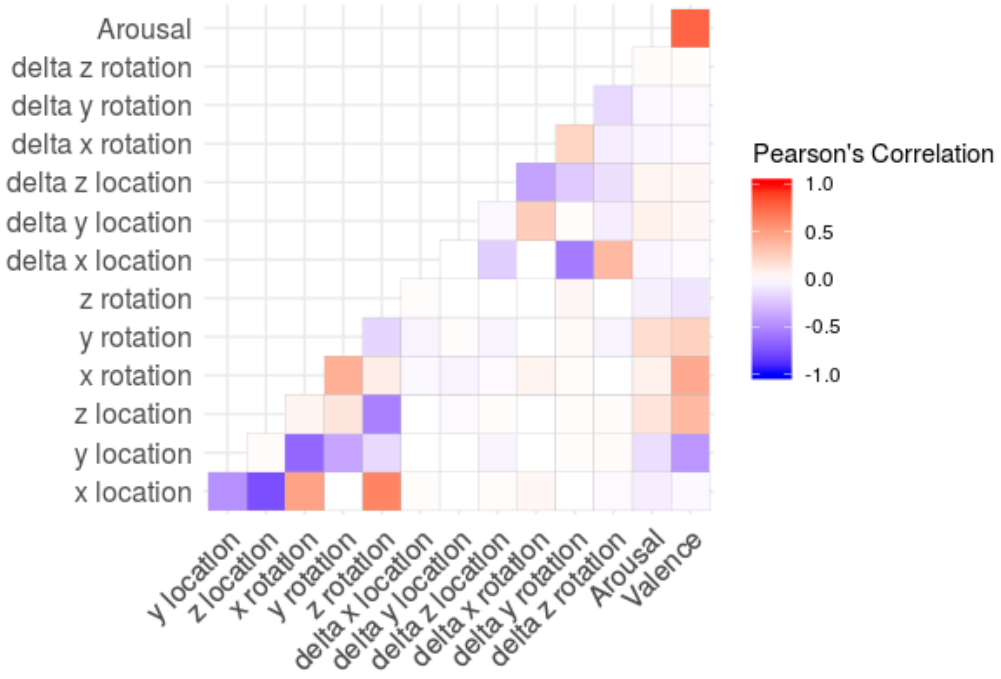
Further features were gathered from the pupil diameter and head rotation and location x, y, z coordinate LLDs in an effort to more fully search these modalities for affective information. For these mid-level features, which are so called because they precede further high-level functional feature extraction, a time-frequency, or more correctly, time-scale, representation was gathered using the discrete wavelet transform.

To measure the frequency components of a signal varying in time one could use the short-time Fourier transform of the signal. This involves applying the fast Fourier transform at various short time windows over a signal. However, this presents users with a dilemma for window selection. Smaller time windows (number of frames) achieve good time resolution but lower frequency resolution while larger time windows achieve better frequency resolution at the cost of less time localisation. This is known as the Gabor (or Heisenberg) uncertainty principle [155]. Wavelet analysis provides an attractive alternative to the short-time Fourier transform where the frequency contents of a signal need to be estimated in time. Wavelets, short and finite oscillating waves with zero mean, have a time-frequency resolution that changes during analysis (see Figure 4.9) and can provide alternative bases to Fourier decomposition. Some advantages of wavelets include more efficient frequency support and the potential to localise aperiodic, or singular, frequency events [156]. This comes at the cost of not providing true frequency domain representations, while additionally wavelets still have to deal with the Gabor uncertainty principle as can be seen in Figure 4.9. Wavelets were employed for this work as abrupt signal changes were expected that could provide insightful temporal features. Furthermore, frequency resolution in the lower bands was important. For example, wavelet representations from pupils of subjects have been used for sympathetic nervous system activation recognition where good frequency resolution was required for the low frequency components (characteristic $f = 0.47 \pm 0.01$Hz) [61]. Also, from dynamic head motion features up to 15Hz, frequencies up to 5Hz were found to be the most significant of the dynamic features for automatic emotion recognition in head movement [65].

Wavelet coefficients can be obtained that are an orthogonal and normalised (or-

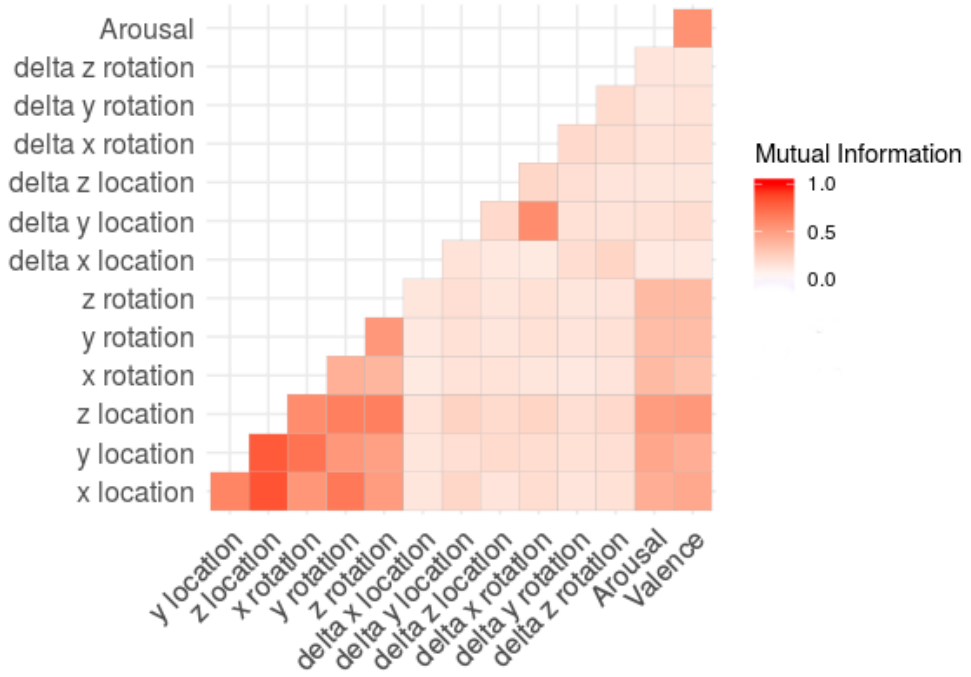


(a)

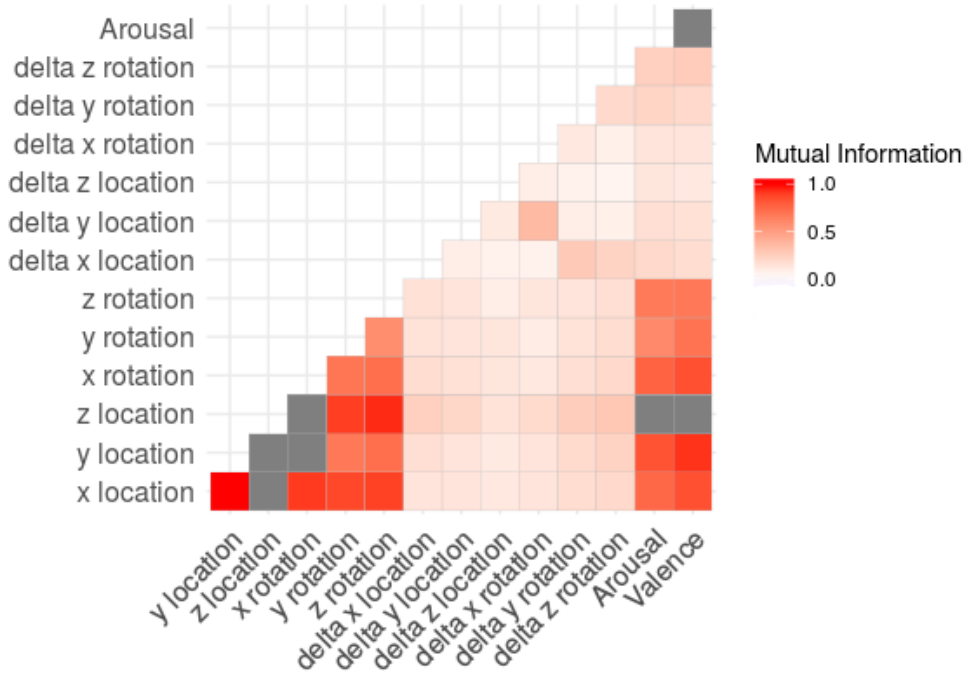


(b)

Figure 4.7: Head pose correlation heatmaps for 8-second moving average of LLDs on (a) RECOLA and (b) SEMAINE.

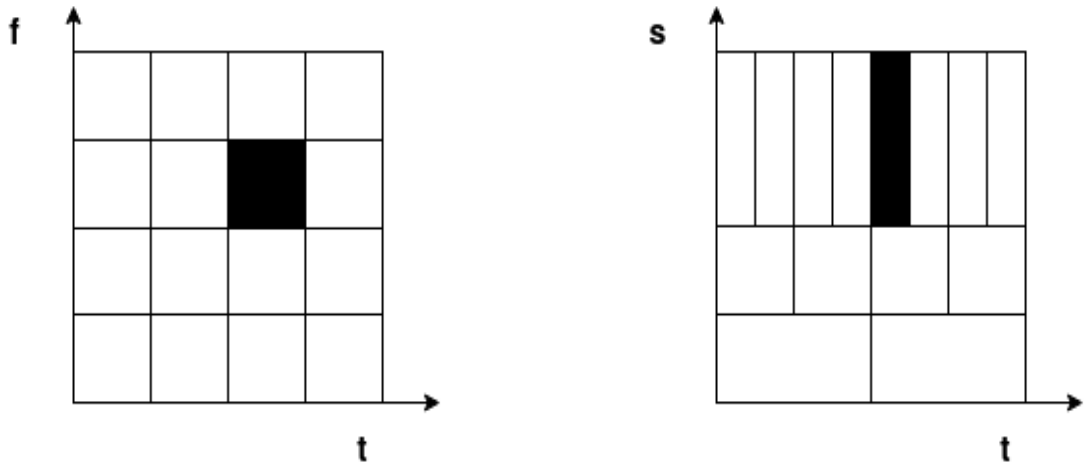


(a)



(b)

Figure 4.8: Head pose MI heatmaps for 8-second moving average of LLDs on (a) RECOLA and (b) SEMAINE. Gray-coloured tiles indicate MI > 1.



(a) (b)

Figure 4.9: (a) Short-time Fourier transform with fixed time and frequency resolution and (b) wavelet decomposition at varying levels of scale (zoom) and time resolution. A frequency of interest is shown in both diagrams with different time and frequency resolutions depending on the analysis method. It can be seen that the wavelet analysis provides better time resolution for high frequency components and better frequency resolution for low frequencies.

thonormal) basis of a square integrable function [157]:

$$\psi_{j,n}(t) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{t - 2^j n}{2^j}\right), \tag{4.2}$$

where t is time, ψ is a mother wavelet function, 2^j is the dilation level, and n the translation. The wavelet transform can be implemented in discrete time by a recursive filter tree, where highpass and lowpass convolution filters provide detail (or wavelet) and approximation (scale) coefficients respectively for signal decomposition. After decomposition into high and low frequency bands and downsampling from the original time-domain signal, further decomposition can occur based on additional filtering and downsampling of lowpass output. Filters implementing the Daubechies wavelet of order 10 [158] were selected for the pupil measurements in this work based on [61]. There is also a scaling function ϕ for this wavelet family, where ϕ replaces ψ in Equation (4.2) to obtain scale coefficients, $\phi_{j,n}(t)$. The Daubechies wavelets [158] are orthogonal and known for having compact support, which is to say, good time localisation. The filters were implemented for 7 levels of signal decomposition on the pupil measurements, which was deemed adequate for characteristic frequency detection associated with affect from this modality [61]. The same wavelet functions and implementation were selected for the head-based time-scale feature representations. However, the head-based signals were only decomposed to 4 levels based on previous work on head-based emotion recognition in which the smal-

lest frequency band assessed was $[0, 1]$Hz [65]. The 4^{th} decomposition level at the 25Hz sampling rate for RECOLA [16], for example, generated wavelet coefficients in the band $[0.78125, 1.5625]$Hz and scale coefficients in the band $[0, 0.78125]$Hz. Following wavelet decomposition, scale and wavelet coefficients for all decomposition levels were taken as mid-level features of the signal to which appropriate functionals were then applied prior to affect learning.

4.2.3 Proposed eye-based feature sets

An initial eye gaze feature set was extracted from three binary [gaze approach, eyes fixated, eye blink/closure] and five numerical [eye blink intensity, x and y gaze angles, and Δx and Δy gaze angles] LLDs described previously. To enhance the LLDs for machine learning, statistics and ratios were employed in the feature engineering process. Further, while the LLDs capture short-term changes (Δs), longer term changes with respect to time were captured using learned regression slopes for each temporal window. Numerous distributional descriptors were also applied, including robust and non-robust measures, and higher-order statistics (distribution skewness, kurtosis). The proposed 79-dimensional eye gaze feature set, GazeVID, is represented by a 79-dimensional feature vector. The full list of features calculated for this set is specified in Table 4.4 (a). Due to the one-sided nature of eye blink intensity (i.e. range $[0.0, 5.0]$), minimum values and some quartile measurements were not extracted from this LLD. Also, of note from Table 4.4 (a) is the absence of minimum values for eye gaze approach time in seconds, due to repeating values for this functional during early experimentation.

In addition to GazeVID other eye-based features were investigated as additions to the initial feature set. Namely, these sets were, eGazeVID, an extended version of GazeVID that includes direct gaze-based features, and EyeVID, which extends eGazeVID further with the inclusion of pupillometry features. These feature sets were more difficult to collect in terms of time, for eGazeVID, and computation, for EyeVID. The proposed 84-dimensional eGazeVID feature set is listed in Table 4.4 (a) and (b) with feature additions to GazeVID highlighted in the (b) section of the table. This feature set extends GazeVID by 5 features, four features based on direct gaze and one feature based on eye blink intensity. The direct gaze features include ratio and time summaries: mean, max, total. Some tested direct gaze summary features were omitted due to value repetition during initial explorations. The 292-dimensional EyeVID feature set contains the calculations listed in Table 4.4 (a), (b), and (c) with feature additions compared to GazeVID/eGazeVID highlighted in section (c) of the table. This set includes static and dynamic-time pupillometry measures in addition to time-frequency dynamic measures in effort to leverage the

Table 4.4: Proposed Affective Eye-based Feature Sets from Video: (a) GazeVID, a 79-dimensional Eye Gaze Feature Set, (b) Features Added to GazeVID to Make eGazeVID, an 84-dimensional Feature Set That Extends GazeVID with Human-knowledge Direct Gaze Annotations, and (c) Features Added to GazeVID and eGazeVID to Make EyeVID, a 292-dimensional Eye-based Feature Set That Includes Both Gaze and Pupillometry Measures.

(a) GazeVID Feature Set

LLDs (8)	Extracted Features
x, y gaze angles, $\Delta x, y$ gaze angles,	min, max, mean, median, quartile 1, quartile 3, skewness, kurtosis, SD, IQR 1-2, IQR 2-3, IQR 1-3, LR slope, LR intercept
eye blink intensity	max, mean, median, quartile 3, SD, IQR 1-2, IQR 2-3, IQR 1-3, LR slope, LR intercept
eyes fixated, eye blink/closure	fixation ratio, fixation time seconds: min, median, mean, max
eye gaze approach	gaze approach ratio, gaze approach time seconds: median, mean, max

(b) Features Added to GazeVID for The eGazeVID Feature Set

LLDs (2)	Extracted Features
eye blink intensity	IQR 1-3
direct gaze	direct gaze ratio, direct gaze time seconds: mean, max, total

(c) Features Added to GazeVID and eGazeVID for The EyeVID Feature Set

LLDs (2) & Mid-level Features (14)	Extracted Features
pupil diameter, Δ pupil diameter	min, max, mean, median, quartile 1, quartile 3, skewness, kurtosis, SD, IQR 1-2, IQR 2-3, IQR 1-3, LR slope, LR intercept
Pupil 10-order Daubechies scale and wavelet coefficients at 7 levels of decomposition	min, max, median, quartile 1, quartile 3, skewness, kurtosis (kurtosis not measured at final decomposition level), SD, IQR 1-2, IQR 2-3, IQR 1-3, RMS, ZCR (ZCR not applied to scale coefficients)

pupil modality for affect prediction in video.

4.2.4 Proposed head-based feature sets

Statistical calculations on the LLDs of head pose from video complete the proposed initial 168-dimensional head pose feature set, PoseVID. The full list of features calculated for this set is specified in Table 4.5 (a). Calculations on both the LLDs and mid-level features of head pose from video complete the proposed increased head pose/motion feature set from video, PoseVID-adv, short for PoseVID advanced. This 768-dimensional feature set is enhanced with more dynamic features approximating frequency representations in time compared to the PoseVID set. This of course does come with the drawback of more requisite computation, however. The features calculated for PoseVID-adv are given in Table 4.5 (a) and (b) with feature additions compared to PoseVID highlighted in section (b) of the table.

Table 4.5: Proposed Affective Head-based Feature Sets from Video: (a) PoseVID, a 168-dimensional Head Pose Feature Set, and (b) Features Added to PoseVID to Make PoseVID-adv, a 768-dimensional Feature Set That Extends PoseVID with Time-frequency Representation Features

(a) PoseVID Feature Set

LLDs (12)	Extracted Features
head location x, y, z Δ head location x, y, z head rotation x, y, z Δ head rotation x, y, z	min, max, mean, median, quartile 1, quartile 3, skewness, kurtosis, SD, IQR 1-2, IQR 2-3, IQR 1-3, LR slope, LR intercept

(b) Features Added to PoseVID for The PoseVID-adv Feature Set

mid-level features (48)	Extracted Features
Head location and rotation x, y, z 10-order Daubechies scale and wavelet coefficients at 4 levels of decomposition	min, max, median, quartile 1, quartile 3, skewness, kurtosis, SD, IQR 1-2, IQR 2-3, IQR 1-3, RMS, ZCR, (ZCR not applied to scale coefficients)

4.3 Unimodal affect prediction experiment design

Motivated by state-of-the-art continuous affect prediction results [12]–[14], this section presents evaluations of deep learning for continuous affect prediction using speech-, eye-, head-, and face-based features from audio-video. During each experimental stage, the method that achieved the best validation set CCC was used for the next experimental stage. Particular experimental steps are further described in the following sections.

4.3.1 Feature extraction temporal window

Features for input to the DNN were extracted using 4, 6 and 8 second temporal windows, moved forward at a rate of 1 frame per interval (refer to Figure 4.4 for this temporal window/chunking method). Each temporal window was tested using speech, head pose (PoseVID), and eye gaze (GazeVID) features, as input to a DNN for arousal and valence training and evaluation (validation). This provided a majority vote as to what temporal window should perform well for further experimental evaluations. The 88-dimensional eGeMAPS [21] speech feature set was gathered from audio data using openSMILE [113]. The proposed head- and eye-based feature sets, were gathered using OpenFace [92] and software developed for this work for LLD, mid-level and functionals features extraction. Temporal feature windows are denoted W_s for the remainder of this work, where s indicates window size in seconds for a window W per interval.

The face feature set used was only extracted and evaluated after the majority vote on the temporal window to be used. The extraction of face features involved functionals extraction from real-valued intensities of AUs in the range $[0.0, 5.0]$ provided from OpenFace [92] for a given temporal window. These LLDs are AU01, AU02,

AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26 and AU45 intensities. The functionals applied included min, quartile 1, median, mean, quartile 3, max, SD, skewness, kurtosis, IQR 1-2, IQR 2-3, IQR 1-3, LR intercept and slope (with time as the independent variable), which resulted in a 170-dimensional feature vector contribution from the face.

4.3.2 Arousal and valence gold standard backward time-shift

Gold standard annotations provided with the RECOLA [16] and SEMAINE [11] corpora were shifted back-in-time to account for annotator rating time delay [3], [13], [90]. The gold standard backward time-shift sizes evaluated ranged from 0 to 4.4 seconds in steps of 0.2 seconds. These are referred to as D_s for the remainder of this work, where s indicates the delay D in seconds applied to gold standard annotations prior to concatenation with input features. The speech, PoseVID, GazeVID, and face feature sets were used at this experimental stage so that modality-wise gold standard backward time-shifts could be found.

4.3.3 Feature selection

Two supervised MI-based feature selection approaches were used to identify a good subset of features from each modality or feature set for affect prediction. MI provides a nonlinear dependency measure between variables, which was a motivation for selecting these approaches in the work presented here using nonlinear machine prediction models. Specifically, a simpler, and less aggressive (assumed, in terms of feature reduction), filter-based MI technique and a more aggressive mRMR [132] technique were used for feature selection. For the filter-based selection, MI is estimated between the features and arousal, and the features and valence on training set using the same method as Section 4.2.1.1. Features with MI less than a defined threshold close to zero were removed. These features were categorised as being independent of arousal or valence due to the lack of shared information and, therefore, poor predictors. The mutual information thresholds evaluated in the experiments were 0.15 nats and 0.2 nats.

The aforementioned filter-based technique is beneficial for its simplicity, however, it does not consider the role of feature-with-feature dependencies for feature selection. Ideally, input features would have low redundancy (low feature-with-feature dependency) and high feature-with-target dependency or relevance. The mRMR [132] algorithm seeks to provide this by incorporating a MI score for each of feature-with-target relevance and feature-with-feature redundancy. The mRMR criterion

is defined as $\max(D - R)$ where D is a feature-with-target relevance and R is a feature-with-feature redundancy. In the experiments, the solution sizes for mRMR features were set to equal half of the original feature vector size to aggressively reduce the feature space. The mRMRe package [159] was used for classical (i.e. not ensemble) mRMR [132] estimation, where, for the continuous variables, MI was estimated based on a correlation measure:

$$I(\mathbf{x}, \mathbf{y}) = -\frac{1}{2}\ln(1 - r(\mathbf{x}, \mathbf{y})^2), \tag{4.3}$$

where r represents the Pearson correlation between $\mathbf{x}$ and $\mathbf{y}$. For the feature selection techniques employed, feature subsets for arousal or valence were first learned and selected on the training set. Following this, the learned feature subsets were selected on the validation partition. At this experimental stage, all the proposed head- and eye-based feature sets were used in addition to speech and face. This experimental step formed the final unimodal deep learning evaluations for the experiment.

4.3.4 Model selection and evaluation

DNN model selection for the experiments was based on the neural network model (and method) that achieved the highest validation set CCC for each modality. In cases where the CCC of two or more DNN models for a modality were equal during experimental evaluation, a second metric, mean-squared error, was considered for model selection.

The CCC scores achieved by the final DNN models for each modality were compared against each other and with the minimum practical baseline scores on the experimental corpora for both arousal and valence. These practical baseline scores are the average annotator CCC values shown in Table 3.2. These CCC scores are 0.293 for arousal and 0.411 for valence on the RECOLA [16] validation set and 0.384 for arousal and 0.684 for valence on the SEMAINE [11] validation set.

4.4 Unimodal affect prediction results and discussion

This section presents the results obtained from the unimodal feature extraction temporal window, gold standard backward time-shift and feature selection experiments involving the proposed head- and eye-based feature sets. Discussion of the effects of these affect learning parameters on arousal and valence prediction, and feature set performances and comparisons against other modalities is included.

4.4.1 Feature extraction temporal windows

Figure 4.10 shows the validation set results achieved on the experimental corpora for both arousal and valence dimensions for the DNNs under different window size (W_s) conditions for speech, head pose (PoseVID) and eye gaze (GazeVID) input. For this experimental stage, the random seed was set at 1787452436. This random seed initialiser was found to work well on initial experiments using speech and was used for the remainder of the experiments in this dissertation. In general, the results indicate that changing the W_s parameter resulted in a change in validation set performance, which is in agreement with the literature [3], [12], [13]. Furthermore, the results suggest that increasing the analysis window size improved affect prediction performance. Over these experiments, increasing the window size by 2 seconds generally resulted in increased affect prediction performance across both corpora and affect dimensions. These results agree with the earlier LLD exploratory analysis in Sections 4.2.1.1 and 4.2.1.2. However, arousal in eye gaze on RECOLA [16] and valence in speech on SEMAINE [11] was best predicted on the validation sets with the W_s parameter set to 4 seconds. These results could be outliers as they do not agree with the consensus of improved performance with widening temporal windows. For the majority vote, setting the W_s parameter to 8 seconds won for this experimental stage and this parameter will be used for the rest of the experiments that follow. This window parameter often provided the best performance unanimously, across the modalities, and, at the very least, it performed best for two out of three of the modalities in question.

Of note from these experiments are the performances of speech, which was the best-performing modality overall across the corpora for arousal prediction, and head pose, the best valence prediction modality on SEMAINE [11]. These results echo how well speech is known to perform for arousal prediction and display potential for valence recognition in head pose, particularly for human-to-agent (agent played by a human) conversation. The results also showed head pose to perform well for arousal prediction from the visual features. This is perhaps because speech, which is known to perform well for arousal prediction, shares affective information with head pose [25], [26]. These results provide early evidence that considering head pose can be of benefit for unimodal affect prediction. They also demonstrate the importance of selecting an appropriate temporal window for feature extraction.

4.4.2 Gold standard backward time-shift

The results from the gold standard time-shift (D_s) experiments are given in Figure 4.11. Top performers from these experiments as measured by validation set CCC on RECOLA [16] were speech for arousal, a CCC of 0.741, and face for valence

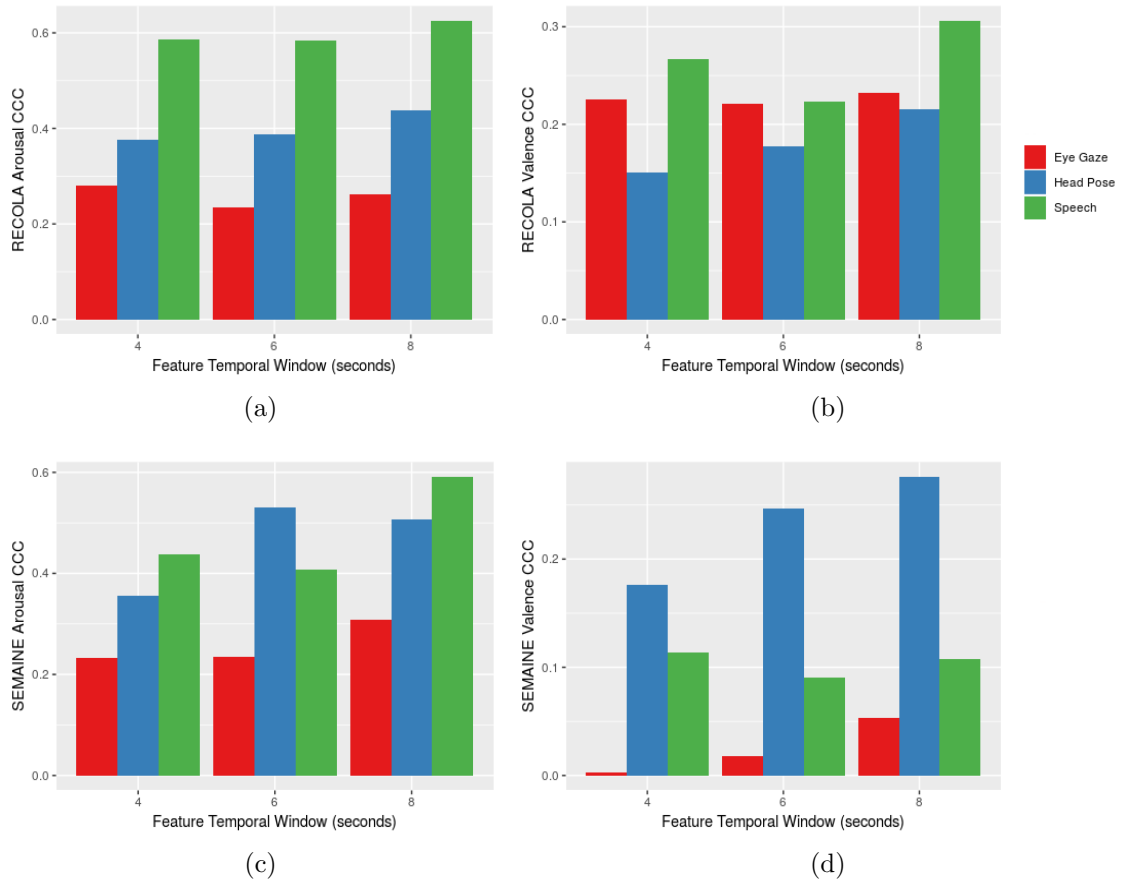


Figure 4.10: Speech, head pose (PoseVID) and eye gaze (GazeVID) DNN input prediction validation set CCC scores for RECOLA: (a) arousal and (b) valence, and SEMAINE: (c) arousal and (d) valence, under different feature temporal window conditions. Temporal window sizes (W_s) of 4, 6 and 8 seconds were evaluated for each modality.

prediction, a CCC of 0.519, using D_s values of 3.2 and 3.4 seconds respectively. Top performers on SEMAINE [11] were speech for arousal, a CCC of 0.680, and head pose for valence, a CCC of 0.289, using D_s values of 3.0 and 0.6 seconds respectively.

The graphs in Figure 4.11 indicate that there was an increase in performance followed by a drop-off when increasing the D_s parameter in modalities except the speech valence condition on RECOLA [16]. This consistency, in general, of improved performance across the corpora when the gold standard is delayed to compensate for annotator reaction time-lag is to be expected [3], [129], [131]. Another similarity across the corpora was that speech always performed best for arousal prediction while the head pose features always performed second-best overall and best from the visual features for arousal prediction. Similar to Section 4.4.1, the head pose features were demonstrated as the best visual descriptor of arousal. They now appear superior to both eye gaze and face features, respectively, for arousal prediction from the visual domain. Some differences noted across the corpora at this exper-

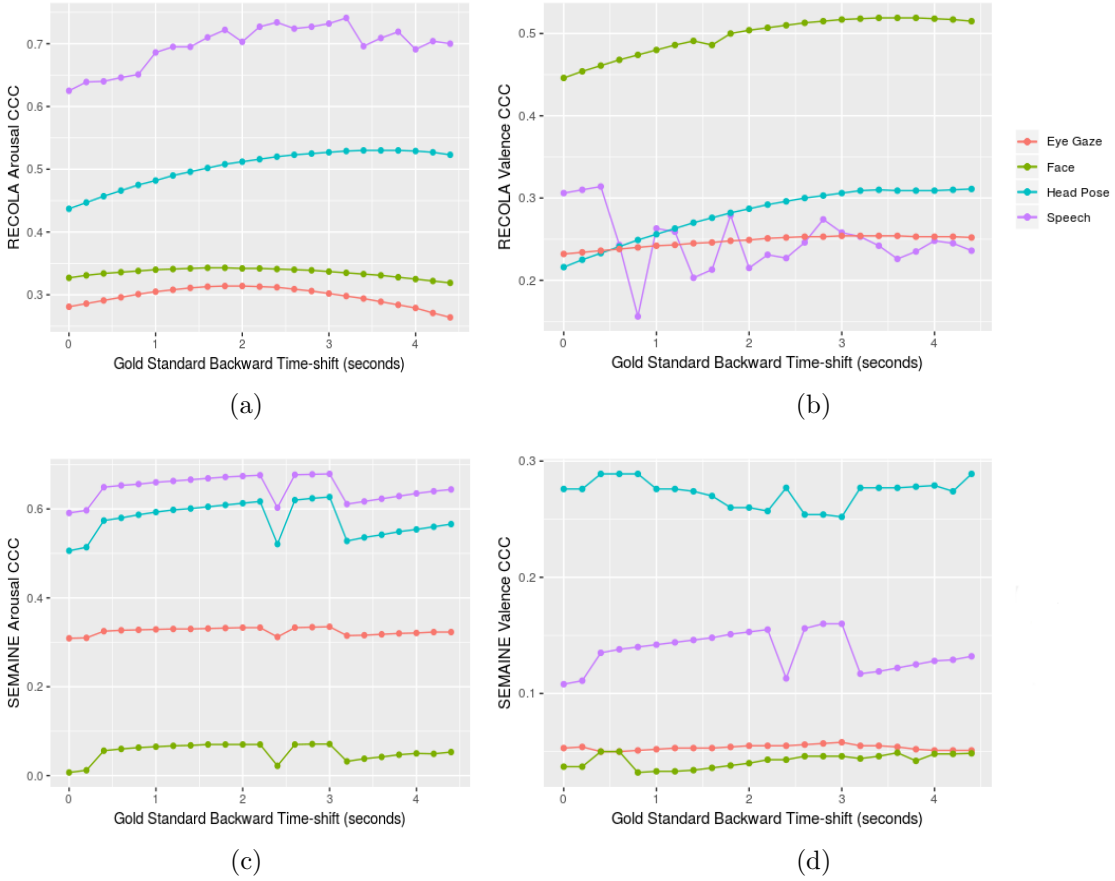


Figure 4.11: Arousal (a) and valence (b) validation set CCC scores under different gold standard backward time-shift conditions. The time shifts (D_s) evaluated ranged from 0 (not applied) to 4.4 seconds, altered in steps of 0.2 seconds.

imental stage were that arousal and valence prediction performance was worse on SEMAINE [11] compared with RECOLA [16] and the difference in the best valence prediction modality. On RECOLA [16], face features provided the best valence prediction performance while on SEMAINE [11] the head pose features were best for valence prediction. This could mean that across social situations, for example, RECOLA [16] involves dyadic problem solving while SEMAINE [11] involves human-to-agent conversation, different modalities may be more important than others to predict valence. Also, the RECOLA [16] subjects were observed gazing down at their desk/notes much more often that the SEMAINE [11] subjects who were only engaging in conversation. It is possible that the task-oriented downward gazing, and therefore head tilt, facilitated less emphasis being placed on head cues as an affective signal by the perceivers/annotators on RECOLA [16].

Table 4.6: DNN Continuous Affect Prediction CCC Results on the (a) RECOLA and (b) SEMAINE Validation Sets for the Best Performing Feature Selection (FS) Method Evaluated for Each Modality (Note: Estimated Group-of-humans Baseline Validation Set CCC Scores: RECOLA Arousal = 0.293, Valence = 0.411; SEMAINE Arousal = 0.384, Valence = 0.684)

(a) RECOLA

	Arousal			Valence		
Feature Set	CCC	*N* Features	FS	CCC	*N* Features	FS
Speech	**.741**	**88**	**N/A**	.370	59	MI .200
PoseVID	.549	121	MI .200	.312	136	MI .150
PoseVID-adv	.572	768	N/A	.341	768	N/A
GazeVID	.340	44	MI .150	.260	49	MI .200
eGazeVID	.378	65	MI .150	.285	42	mRMR
EyeVID	.339	146	mRMR	.276	113	MI .200
Face	.426	85	mRMR	**.560**	**66**	**MI .200**

(b) SEMAINE

	Arousal			Valence		
Feature Set	CCC	*N* Features	FS	CCC	*N* Features	FS
Speech	**.736**	**44**	**mRMR**	.160	88	N/A
PoseVID	.638	159	MI .150	**.289**	**168**	**N/A**
PoseVID-adv	.671	768	N/A	.285	694	MI .150
GazeVID	.417	40	mRMR	.058	79	N/A
eGazeVID	.330	42	mRMR	.100	74	MI .150
EyeVID	.198	260	MI .150	.219	225	MI .200
Face	.101	168	MI .150	.160	163	MI .200

4.4.3 Feature selection

The best results obtained as measured by validation set CCC for the feature selection experiments can be observed in Table 4.6. As shown in the table, no feature reduction resulted in prediction performance improvements for two modality/affect evaluations on RECOLA [16] and four modalities on SEMAINE [11]. Generally, however, the feature sets had their sizes reduced along with performance increases on the corpora validation sets, showing the employed methods to be effective. For arousal, the mRMR [132] algorithm was often effective at providing the best feature subset and performance for the modalities and feature sets evaluated. The larger feature reductions produced by mRMR [132] is also attractive. For example, the best performing modality for arousal on SEMAINE [11] was speech where a CCC of 0.736 was achieved along with a feature set size reduction of 50%. For valence, mRMR [132] was less successful than the MI filter-based technique for feature selection. This could be due to a good subset in valence feature space being harder to find, as has been previously suggested [160], rendering the more aggressive feature selection technique less fruitful.

For arousal prediction speech was once again a clear winner, in terms of performance, across both corpora. The eGeMAPS [21] speech feature set used provides computation and potential interpretability benefits due to its relatively small size in addition to the performance benefits that have been demonstrated here. Head-based features were once again second-best overall for arousal prediction and best

for arousal prediction from the visual feature sets. The PoseVID-adv feature set performed best from the head-based sets. It provided relative performance improvements above PoseVID of 4.19% on RECOLA [16] and 5.17% on SEMAINE [11]. These performance increases were provided by a larger feature vector from the head, however, where over six hundred extra features were used by PoseVID-adv to provide these performance increases. Relative to speech, the PoseVID-adv feature set performed 22.81% worse on RECOLA [16] and 8.83% worse on SEMAINE [11] for arousal prediction in terms of CCC. Another notable performance was that of eye gaze on SEMAINE [11] where a validation set CCC of 0.417 was achieved with the GazeVID set. Curiously, the GazeVID feature set, the smallest of the eye-based sets, was the best-performing from the eye feature sets on this corpus and affect and it performed better than face features as well. Unfortunately, GazeVID, or any eye-based feature set, did not consistently provide better performance for arousal prediction across the corpora. The face features outperformed the eye features for arousal prediction on RECOLA [16]. The eye-based feature sets did not perform within a comparable range to that of speech on either experimental corpus.

There was no clear leader for unimodal valence prediction, across the experimental corpora. The face feature set performed best for valence prediction on RECOLA [16] and there were no other modalities that were within a comparable range of the face system's performance. Head pose, PoseVID specifically, performed best for valence prediction on SEMAINE [11] while the larger PoseVID-adv feature set performed just 1.38% worse relative to the PoseVID set's performance for valence prediction. The head-based performances may have been caused by the annotators for SEMAINE [11] placing more emphasis on head gestures when forming their valence ratings. Also, as mentioned in the previous section, there is the possibility that task-oriented downward gazing and head pose may de-emphasise the affective signalling component of head pose on RECOLA [16]. Interestingly, for valence prediction from the eye-based sets, the largest eye-based set, EyeVID, performed best for eye-based valence recognition on SEMAINE [11] but not on RECOLA [16]. This may have occurred due to the more professional lighting and recording set-up used in SEMAINE [11]. Unfortunately, if this prediction is true, this will limit the use of the EyeVID feature set to very experimentally controlled or professional recording settings.

The automatic prediction models were able to outperform the group-of-humans baseline prediction estimations regularly for arousal across the experimental corpora. For example, all the systems outperformed this baseline CCC of 0.293 on RECOLA [16] while 4 of 7 (speech, eye gaze and both head pose) systems outperformed the baseline CCC value of 0.384 on SEMAINE [11]. The automatic prediction of valence appeared much more difficult in comparison. Only the face feature

input system outperformed the human performance CCC estimate of 0.411 on the RECOLA [16] validation set. No automatic system could outperform the group-of-humans benchmark for valence on SEMAINE [11]. The phenomenon of automatic prediction systems not being able to match or exceed human performance estimates for valence on SEMAINE [11] is not new [125]. This shows how difficult it is to learn the valence in the subjects of the SEMAINE [11] corpus. A reason for this difficulty may include the culture variability in SEMAINE [11], for example, Irish, North American, and mainland Europe cultures are present in the subset taken for this work. Also, there was a smaller subset taken for the SEMAINE [11] training set compared to that of RECOLA [16] due to the amount of direct gaze annotations available for SEMAINE [11]. This potentially exacerbated the cultural variability issue, where there were a smaller number of training patterns to learn from.

4.4.4 General discussion

Head-based continuous affect prediction has achieved results comparable to speech in unimodal settings in the past [72]. The results in this chapter suggest that head pose features are second-best when compared with speech for unimodal arousal prediction. The head-based feature set, PoseVID-adv, performed notably well for arousal prediction, however. Also, speech and head cues are related for affective signalling [25], [26]. Based on this, and the performance obtained by head-based cues in the experiments, there may be potential to exploit the head modality in different ways to benefit affect prediction systems. For example, head pose features might be of benefit to predict certain speech-based features when the audio signal is itself overly noisy. Alternatively, head-based predictions could be employed as a secondary prediction method in multimodal settings in cases of severely poor audio conditions. The eye-based feature sets proposed appear inappropriate, performance-wise, for unimodal arousal prediction. It is still believed, however, that the eye features can be shown to be useful in multimodal settings.

There was no clear best modality across the experimental corpora for valence prediction. Valence is more subject-dependent [139] and it is possible that features that generally perform well, across subjects, cultures and social situations, are harder to find for this affect. Human expression analysis is also generally best approached by using multimodal approaches [24]. For the more difficult-to-predict valence dimension, the experimental results in this chapter verify this as the valence prediction results were considerably worse than that of arousal.

Some limitations of the experimental results presented in this chapter include the unimodal approach taken, which does not consider the complimentarity of the proposed features with speech and facial features. Furthermore, there may be ad-

ditional cross-modal features from speech, head, eye, and face modalities that can provide further information in multimodal settings. Finally, the valence prediction results in the experiments were poorer than those of arousal. While this is to be expected [76], further effort toward increasing the performance of valence prediction should be explored such as investigating the correlation between arousal and valence for modelling [35].

4.5 Conclusion

This chapter set out to propose feature sets from head- and eye-based cues gathered from video and to answer the research question:

How well do head- and eye-based features perform compared with speech and facial features for unimodal continuous affect prediction?

From the experiments conducted, only the head-based features were consistently shown as suitable for unimodal arousal prediction. The PoseVID-adv feature set performed best for arousal prediction from the head-based sets proposed. In comparison to unimodal speech-based performance, the best arousal prediction modality overall, the head-based features performed less well. Specifically, head-based arousal prediction CCC scores were 22.81% less on RECOLA [16] and 8.83% less on SEMAINE [11], relative to speech arousal CCCs. The PoseVID-adv feature set also outperformed estimated human performance baseline CCC values across both experimental corpora. The head-based feature set, PoseVID-adv, was the best visual feature set, compared to the eye or face sets, for arousal prediction. For valence prediction, the experimental results showed face features to perform best on RECOLA [16] and the head feature set, PoseVID, to perform best on SEMAINE [11]. No other modalities were comparable, in terms of valence performance on these respective corpora. Only the face feature set performance outperformed the estimated human performance baseline, however. It is therefore concluded from the experiments that head-based features are suitable for unimodal arousal prediction and -15.82% relative CCC performance can be obtained from this modality compared with unimodal speech. It should be noted that many more features (hundreds) are also required from head pose to obtain such performance. From the valence experimentation carried out, it is concluded that head- and eye-based features are not suitable for unimodal valence prediction. This is because, even while the head-based features performed best in the face-to-face conversation situation of SEMAINE [11], they did not outperform the estimated human performance measure.

The next chapter presents results of multimodal affect prediction evaluations employing the best-performing head- and eye-based feature sets proposed in this chapter. Feature fusion and cross-modal feature engineering is investigated to optimally leverage the complimentarity of the proposed features with speech and face feature sets. Teacher-forced learning was also investigated to improve valence modelling methodology by exploiting the correlation between arousal and valence.

Chapter 5

Multimodal and Teacher-forced Learning Experiments

5.1 Introduction

Head-based features were shown to provide good performance for unimodal arousal prediction from the visual features evaluated in Chapter 4. Due to the unimodal methodology employed in Chapter 4, the complimentarity and/or redundancy of modalities and features or whether there exist any beneficial cross-modal affective features (e.g. holding one's head high and speaking with joy) could not be considered. Therefore, the focus of this chapter is multimodal continuous affect prediction where the fusion of speech, head, eye, and face features are considered for this task. The identified research question for this chapter is

How much of an improvement can head- and eye-based features provide when included in multimodal continuous affect prediction systems?

It should be noted that, while face features are included for completeness in this work, the main goal of this research was to improve upon speech-based performance. Performance differences for multimodal systems were measured against the best of either unimodal speech or bimodal speech and face performances for each affect dimension. This was done to provide a holistic view of the expected improvement when using the proposed head- and eye-based feature sets as one might use all available modalities in an audio-video stream to improve system performance. Both early feature fusion, hereafter, feature fusion, and model fusion were evaluated as methods of combining the modality feature sets in question. Also, while it has long been noted that an open research opportunity involves investigating cross-modal feature interactions [35], this has only been done implicitly (i.e. within algorithms).

It is therefore warranted that the space of speech-, head-, eye- and face-based affective features be searched further by extracting signals based on some combination of these cues toward uncovering cross-modal interaction features.

Finally, teacher-forced learning with multi-stage regression (TFL-MSR) was used to leverage the correlation between arousal and valence and put arousal annotations (i.e. from the teacher/annotators) to further use for valence modelling. TFL-MSR is not tied to a particular fusion method, like strength modelling [15] but unlike OA fusion, for example. A difference between TFL-MSR and strength modelling [15], where preliminary predictions are used to strengthen subsequent models, however, is that TFL-MSR only uses arousal prediction values on the test set to improve valence prediction. The arousal training inputs for TFL-MSR valence prediction are the actual gold standard arousal annotations, hence the term teacher-forced learning.

Affect prediction experiments were carried out on the RECOLA [16] and SEMAINE [11] corpora to evaluate the feature combinations and other experimental methods used. The results showed that multimodal systems outperformed unimodal ones, with model fusion performing best overall on the validation sets. There were very few cross-modal interaction features found and TFL-MSR was shown to improve valence prediction on RECOLA [16] but not SEMAINE [11]. The affect prediction experiments were followed by a feature-with-target relationship analysis that demonstrated the strongest relationships with arousal and valence for the head and eye features selected for the final models.

5.2 Multimodal and teacher-forced learning experiment design

This experiment evaluated multimodal input DNN-based models for continuous affect prediction. In addition to feature fusion and model fusion, cross-modal interaction features, an algorithm sensitivity analysis and TFL-MSR were also investigated in the experiments. The experimental steps are shown in Figure 5.1. The techniques used are further discussed in the sections to follow.

5.2.1 Feature extraction, gold standard backward time-shift, and feature selection

The eGeMAPS [21] speech feature set was gathered from audio data using openSMILE [113] as the core speech feature set for the experiments. However, the ComParE 2013 [31] LLDs and their first order derivatives were also extracted using openSMILE

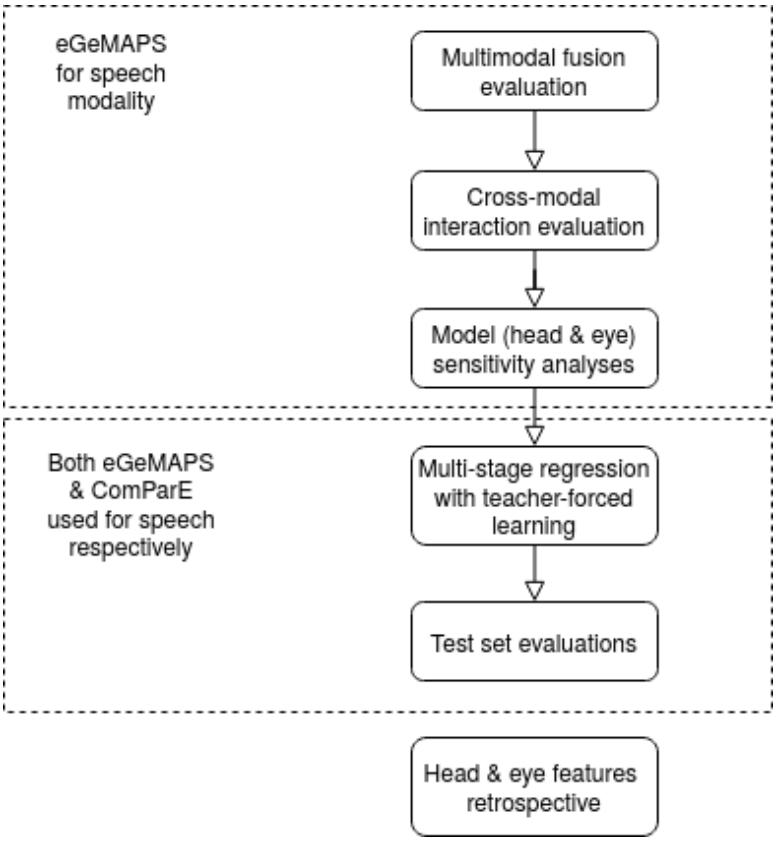


Figure 5.1: Experimental steps. For (early) feature fusion, gold standard backward time-shift and feature selection was carried out. Following this, the feature fusion evaluation, using feature fusion and model fusion were performed where the best modality feature sets and selection techniques from Chapter 4 were used for model fusion. This was followed by cross-modal feature generation, and these features were combined with feature fusion feature vectors for evaluation. The model (head & eye) sensitivity analyses were then performed by way of feature group screening-based sensitivity analysis. The affect prediction experiments culminated with the multi-stage regression using teacher-forced arousal features and final model evaluations on the test set. After the main experiments (outside the broken-line boxes) the head- and eye-based features' relationships with arousal and valence was evaluated for the features from these modalities that were selected for the final systems.

[113]. From these LLDs, min, mean, max, range and SD functionals were extracted. This was done to allow fair comparison against the best performing arousal prediction system from the literature [12]. These features were used instead of the eGeMAPS [21] features for the speech contribution in a small set of final validation set experiments. While no automatic feature selection was applied to the ComParE 2013 [31] LLD-based functionals, six of the functionals had to be manually removed from the set due to the occurrence of repeating values. The resulting feature vector was of size 644.

The speech, head, eye and face features for input to the DNN were extracted using an 8-second analysis window, moved forward at a rate of 1 frame per inter-

val. This window size generally performed best for affect learning and prediction in Chapter 4 (Section 4.4.1) and was used here for all modalities and modality combinations. For this experiment, only windows that contained all modalities were used for prediction to allow accurate assessment of modality contributions when they are all considered together. The resulting training, validation and test set sizes were 57,190, 57,906, and 50,382 respectively for RECOLA [16] and 46,096, 46,537 and 56,286 respectively for SEMAINE [11].

The gold standard backward time-shifts applied for unimodal speech-, head-, eye- and face-based feature sets for model fusion were the same as those used in Section 4.4.2. Additionally, the feature set and feature selection method used for these unimodal contributions to model fusion were the same as those from the unimodal evaluations of Chapter 4 (Section 4.4.3). For example, PoseVID-adv with no feature selection was always used, across the corpora, to model arousal from head-based features and EyeVID with MI feature selection was used to model valence from eye-based features on SEMAINE [11]. Gold standard backward time-shifts were estimated for feature fusion based on maximising DNN CCC performance on the validation set for 23 values of D_s parameters tested, i.e. $D_s = 0, 0.2, 0.4, ...4.4$. Feature selection for feature fusion was carried out using the same mutual information-based methods presented in Chapter 4 for each modality combination investigated.

5.2.2 Feature fusion

Both feature fusion and model fusion were employed for the experiments. Due to the focus on head- and eye-based feature research in this work, the following feature combinations were evaluated: speech & head, speech & eye, speech & face, head & eye, speech & head & eye, face & head & eye and all modalities. This allowed comparisons of the proposed features against the face features, both with and without speech features, and combined with each other in multimodal systems. For feature fusion, sometimes referred to as early feature fusion, the row-wise concatenation of features from each modality into one, larger feature vector was performed prior to DNN training. This fusion method allows for numerous interactions of features inside a DNN.

Model fusion was achieved using the same method as [13], therefore, predictions from each individual modality were fused after convolving them with a Gaussian filter with a window of size 120 frames. The σ parameter of the Gaussian, which controls finite Gaussian support size, was set to 2.5. This fusion process is shown in Figure 5.2. The filter tail values were not used for preparing the training and validation examples. Therefore, 60 samples from the start and end of this set were removed, resulting in 120 fewer examples in the experimental corpora validation

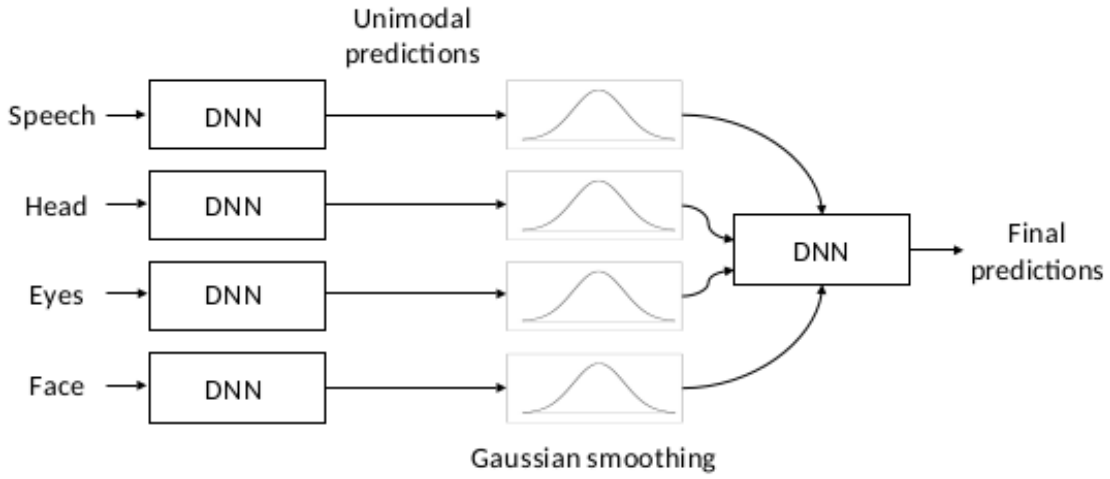


Figure 5.2: Gaussian-smoothed DNN model fusion.

sets for model training and evaluation. This gave more confidence in the final training examples because predictions were only provided where the filter was fully applied to the initial prediction signal. This came at the cost of increased delay (no predictions for first 60 samples) and short-stopping (no predictions for last 60 samples) for filtered predictions. Smoothed predictions were then passed as training examples to another DNN with the same topology. The same training methodology was used for the fusion DNN as for the individual modality DNNs. Unimodal DNN predictions on the validation set were split randomly 50-50 (training-validation %) in order to try to provide a balance between training data and validation data. This randomisation was carried out using the `set.seed()` function from the R base package with a seed value argument of 1. The partitioning aimed to achieve enough training fit while additionally providing good validation set performance estimation. Validation set CCC calculations for the model fusion performances were always based on the unimodal speech gold standard time-shift parameter (Section 4.4.2). This was done as speech is a key modality to be improved upon when predicting affect using multimodal methods.

5.2.3 Cross-modal feature generation and evaluation

Following the initial fusion evaluation of the feature sets, interactions across feature modalities were investigated to identify potentially salient predictors of affect. This was based on results where head pose/gesture changed emotion perception during utterances [25], [26], [65], pupils changed after utterances during a memory task [29], and head pose assisted facial emotion recognition [24]. To search the cross-modal interaction feature space, addition, subtraction, multiplication and division interactions were assessed using autoML. Specifically, a proprietary tool, h2o driverlessAI (version 1.7) was used, where a generalised linear model (GLM) was used for predic-

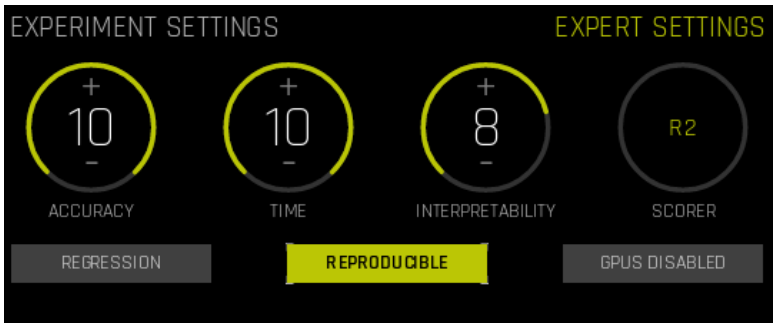


Figure 5.3: h2o driverlessAI basic experimental settings.

tion to evaluate generated features. A GLM extends linear models to handle error distributions other than Gaussian, unifying a number of regression approaches. The identity link was used in this work, which is the LR case of a GLM. The results taken from driverlessAI were the generated interaction features and not the models or predictions. While it is accepted that multimodal modelling is beneficial for affect prediction, what feature interactions are salient for prediction is often unknown and hidden within the model. The purpose of this investigation was to identify some of these features.

For this feature generation method, all feature vectors from each modality were combined using feature fusion with no feature selection. This was done to search the feature space as thoroughly as possible for potential cross-modal feature interactions. The driverlessAI software was used with settings shown in Figure 5.3. The R^2 scorer used is defined as

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu_y)^2} \tag{5.1}$$

and it provides a measure of explained variance of y by $\hat{y}$ that can also be interpreted as the square of the Pearson's r between y and $\hat{y}$. This scorer selection was made as it was the best match for r, which has previously been used for continuous affect prediction performance measurement, from the scorers available in driverlessAI. The random seed value for both arousal and valence experiments on both corpora was 1,234. The driverlessAI expert settings for the experiments included feature engineering effort (10), target transformation (identity), enable target encoding (disabled), while the only enabled learning algorithm was GLM.

Features that were obtained were then evaluated within the core experiment. To achieve this, cross-modal features were merged with the corresponding best-performing arousal and valence feature fusion sets, then training and validation set prediction was performed. For example, the best performing face & head & eye feature fusion system had its best-performing feature vector increased by n features, where n were cross-modal interaction features between a pair from these three mod-

alities. The features were not evaluated with the model fusion approach because each unimodal model for fusion clearly disallows cross-modal feature interactions.

5.2.4 Feature group screening-based sensitivity analysis

Inspired by fractional factorial design [161], a screening-based algorithm sensitivity analysis was performed for head- and eye-based feature group contributions to the prediction systems. This was carried out to provide information on important factors in terms of feature groups for affect prediction from the proposed sets. Another reason for this analysis was to select the highest performing feature sets should feature selection have not provided the best solutions. The screening was carried out using feature group removal, where only one feature group is removed prior to training and evaluation at each step of the analysis. The nine possible head-based feature group removals at this stage were: all head location, all head rotation, head location and rotation x, y, and z respectively, and all simple head location and rotation features (i.e. PoseVID). The eleven possible eye-based feature group removals included: eye gaze approach, direct gaze, x and y gaze angles, eyes fixated, eye closure/blink, eye blink intensity, pupil constriction, pupil dilation, and pupil diameter features (including wavelet features).

Feature groups that resulted in a large relative performance drop were deemed more important for modelling. They contributed notably to the overall model performance. Also, in the event that performance *increases* occurred after feature group removal, these feature groups were deemed poorer predictors and were candidates for removal from the feature sets for the final models. Performance decreases were generally expected at this experimental stage, however, it was also believed that some feature groups were redundant, harming performance. Therefore, feature group removals that provided performance benefit increases of one standard deviation or more above the average CCC for the modality screening were retained for final model evaluations.

5.2.5 Teacher-forced learning with multi-stage regression

It is generally known, and it has been shown in Section 3.2.4.3, that there are correlations between arousal and valence. However, the community has not investigated teacher-forced learning [102, p. 377] as an improved method for affect modelling. Teacher-forced features are traditionally applied by incorporating $t-1$ ground-truth values as additional features for prediction at time t in a RNN. This work instead proposes separating target values by affect dimension instead of time so that arousal knowledge (from annotators) and predictions can be used for learning and predicting valence. The described approach, teacher-forced learning, was investigated and

the results compared against a baseline (i.e. standard regression) and multi-task learning, which has often been used for joint modelling of affect dimensions [12], [120], [138]. The teacher-forced features were employed only for valence learning, the teacher-forced learning stage of this method, in the experiments. That is, the annotations provided for arousal were added as an input feature vector such that a valence feature vector of size N is increased to $N + 1$. This process was carried out for valence training and validation data only as test data annotations must be left unseen. In all cases, paired arousal data used as features had gold standard time-shifts applied that matched the best-performing unimodal arousal systems. For the final test set pass, multi-stage regression was implemented where predictions from the best performing arousal model were first made on the test set. These predictions were then passed as feature input to the final valence model in addition to the traditional input features to complete the proposed TFL-MSR approach. This process was carried out under the assumptions that arousal can be predicted with higher fidelity than valence [31], [76].

TFL-MSR is similar to MTL [120], [138], in that it aims to leverage the correlation of arousal and valence for better feature representation in neural networks. A key difference between TFL-MSR and MTL, however, is that TFL-MSR uses arousal as an additional feature, as opposed to a target like in MTL. This also differentiates TFL-MSR from strength modelling [15] in which preliminary predictions are used to strengthen subsequent models and affect dimensions are treated separately. The TFL-MSR approach to valence prediction uses knowledge gained for arousal prediction toward improved valence prediction on a test set. This was investigated to improve valence prediction performance as it appeared more difficult to predict from the results in Section 4.4.3.

5.2.6 Final model evaluations

A test set pass was carried out for arousal and valence prediction for the highest performing models from the validation experiments to provide unbiased estimates of model performance on each experimental corpus. However, additional test set passes were carried out on the experimental corpora for two reasons. First, both systems incorporating the ComParE 2013 [31] LLD functionals and eGeMAPS were run on the test sets. This was to ensure fair system comparisons with [12], which used ComParE 2013 [31] LLD-based functionals, while also allowing fair comparison with published eGeMAPS test set results [13], [14]. Furthermore, close comparison against [12] is important because the authors used head-based features as part of multimodal input for affect prediction. Second, to provide validity for the TFL-MSR method, a model that used this method and a model that did not were applied to the

test partition where TFL-MSR performed best on the validation set. This provides the best assessment of the improvement of this method above standard regression, that is, on both validation and test data. Before all test set passes, however, the final models were selected based on validation set performance. This provides information detailing the biased and unbiased estimates of model performance on the test set. Final model results were compared to both average annotator CCC and independent results from the literature previously cited.

5.2.7 Post-hoc head- and eye-based feature retrospective

For the feature retrospective, the top 10 performing head- and eye-based features from the experiments were listed in terms of r and MI dependency measures with arousal and valence. These features were compared in terms of r and MI to the best-performing speech and face features in terms of arousal and valence relationships on the RECOLA [16] and SEMAINE [11] training sets. This was done to show, in a model agnostic way, which features from the proposed sets are important descriptors of arousal and valence.

5.3 Multimodal and teacher-forced learning results and discussion

This section presents the results obtained from the multimodal experiments. Comparisons and discussion of the results achieved compared with independent results from the literature and analysis of important feature relationships with arousal and valence are also provided.

5.3.1 Multimodal fusion

The best performing D_s parameter evaluation results for the multimodal feature fusion combinations investigated in this experiment are given in Table 5.1. In general, a lower gold standard time-shift was found for valence than arousal. Interestingly, however, for the fusion of all modalities, the arousal gold standard time-shift was always lower than that of valence across both of the corpora. Assuming that the gold standard annotators used all modalities available to them when forming their ratings, perhaps the fusion of all modalities provides the best estimator of the delay in providing annotations. Of course, different modalities may be more salient contributors to ratings of different affect dimensions at different times. In assessing the performance of feature fusion obtained across affect dimensions and corpora, Table 5.1 shows that the multimodal (all modalities) systems performed best for

valence prediction but not for arousal prediction. The best feature fusion arousal systems used bimodal speech and another modality on both the SEMAINE [11] and RECOLA [16] validation sets. This indicates that, while valence may be more unconsciously recognisable than arousal [76], conscious valence perception and annotation actions that follow may take longer than that of arousal by requiring more information from more modalities.

The results for the best performing feature selection methods for feature fusion and combinations of unimodal models usings model fusion can be observed in Table 5.1. For arousal prediction, it can be seen that model fusion always outperformed feature fusion. For head & eyes and face & head & eyes feature fusion, there is a drop in performance when comparing the Table 5.1 results with the unimodal head results in Table 4.6 for both corpora. In further comparisons of the feature fusion performances with the unimodal results from Table 4.6, it was noted that only feature fusion systems that included face features on RECOLA [16] for valence were able to improve prediction. The model fusion systems largely outperformed the unimodal systems of Chapter 4, however. These results suggest some redundancies across the feature modalities that resulted in poor feature interaction representations or perhaps sub-optimal fusion of the modalities, temporally, with the affect to be predicted.

The best arousal results in Table 5.1, CCC values of 0.833 on RECOLA [16] and 0.859 on SEMAINE [11], show the potential of combining all the employed modalities for prediction. For these cases, features for arousal prediction were integrated at different unimodal D_s parameter values, therefore, this model could take advantage of modality information at different levels of time integration. Other notable performances from the arousal model fusion results was that of speech & head & eyes, on both corpora. On RECOLA [16], this system achieved a CCC of 0.809, a performance 2.88% lower relative to the top performing system along with 86 fewer features compared to the system that used all modalities. On SEMAINE [11], a CCC of 0.831 was achieved by the speech & head & eye system, a performance 3.26% lower relative to the top performing all modalities system along with 169 fewer features.

The valence results in Table 5.1 appear to repeat the outcome of the arousal experimentation, that model fusion is superior to feature fusion. Model fusion provided the best performance from the valence experiments, providing CCC values of 0.694 on RECOLA [16] and 0.843 on SEMAINE [11], with all modalities used as input on both corpora evaluations. From Table 5.1 (b) it is noted that the visual feature fusion systems always outperformed the model fusion systems, however. This shows that, for valence prediction, the visual features can offer complimentary interactions for feature representation inside DNNs for in some situations, specifically, the RE-

Table 5.1: Multimodal Fusion DNN Validation Set Prediction CCC Results on RECOLA for (a) Arousal and (b) Valence, and SEMAINE for (c) Arousal and (d) Valence for the Best Performing Feature Selection (FS) Method and D_s Parameter Evaluated with Resulting Feature Vector Sizes Listed as N Features (Note: Estimated Group-of-humans Baseline Validation Set CCC Scores: RECOLA Arousal = 0.293, Valence = 0.411; SEMAINE Arousal = 0.384, Valence = 0.684)

(a) RECOLA Arousal

	Feature Fusion				Model Fusion	
Modalities	CCC	N Features	FS	D_s	CCC	N Features
Speech & head	.717	679	MI .150	2.6	.796	858
Speech & eyes	.728	86	mRMR	3.2	.787	155
Speech & face	.740	129	mRMR	3.6	.794	175
Head & eyes	.558	549	MI .200	2.8	.573	835
Speech & head & eyes	.733	470	mRMR	3.4	.809	924
Face & head & eyes	.544	1,022	N/A	2.8	.608	921
All modalities	.705	731	MI .200	3.6	**.833**	**1010**

(b) RECOLA Valence

	Feature Fusion				Model Fusion	
Modalities	CCC	N Features	FS	D_s	CCC	N Features
Speech & head	.473	682	MI .150	2.4	.433	836
Speech & eyes	.265	116	MI .200	2.0	.435	110
Speech & face	.545	187	MI .200	3.2	.587	186
Head & eyes	.421	852	N/A	3.2	.337	812
Speech & head & eyes	.533	470	mRMR	2.4	.519	879
Face & head & eyes	.631	1,022	N/A	3.2	.588	931
All modalities	.639	1,110	N/A	3.8	**.694**	**998**

(c) SEMAINE Arousal

	Feature Fusion				Model Fusion	
Modalities	CCC	N Features	FS	D_s	CCC	N Features
Speech & head	.694	783	MI .150	3.0	.818	814
Speech & eyes	.706	mRMR	86	3.0	.785	86
Speech & face	.563	254	MI .150	2.8	.759	214
Head & eyes	.559	765	MI .150	3.0	.772	810
Speech & head & eyes	.705	468	mRMR	3.0	.831	855
Face & head & eyes	.527	933	MI .150	3.0	.807	979
All modalities	.538	1,105	N/A	2.2	**.859**	**1,024**

(d) SEMAINE Valence

	Feature Fusion				Model Fusion	
Modalities	CCC	N Features	FS	D_s	CCC	N Features
Speech & head	.276	246	MI .150	3.0	.524	258
Speech & eyes	.195	330	MI .150	3.0	.680	315
Speech & face	.131	129	mRMR	2.8	.484	253
Head & eyes	.260	460	N/A	2.8	.721	1,395
Speech & head & eyes	.278	548	N/A	1.4	.804	484
Face & head & eyes	.203	568	MI .150	2.8	.756	559
All modalities	.284	718	N/A	3.8	**.843**	**648**

COLA [16] problem solving task in this case. Of further note for the face & head & eyes visual systems in Table 5.1 (b) is that they indicate practical benefit on the validation set using either fusion method on RECOLA [16]. These multimodal systems outperformed the average CCC estimated from a group of human annotators, a CCC score of 0.411. Only the model fusion systems in Table 5.1 (d) were able to outperform the group of human annotators estimate for valence, a CCC score of 0.684, on SEMAINE [11].

Table 5.2: h2o driverlessAI autoML-Generated Feature Interactions Extracted Based on Validation Set Performances for (a) RECOLA Arousal, (b) RECOLA Valence and (c) SEMAINE Arousal

(a) RECOLA Arousal

Features
HNRdBACF_sma3nz_amean / loudness_sma3_amean, db_scale_coeffs_l3_head_rotation_z_min + db_scale_coeffs_l4_head_rotation_x_median

(b) RECOLA Valence

Features
AU12_SD / delta_gaze_angle_x_quartile_3

(b) SEMAINE Arousal

Features
db_scale_coeffs_l1_head_rotation_y_skewness × delta_head_location_z_median

5.3.2 Cross-modal interaction features

The results from the cross-modal interaction feature generation are given in Table 5.2, with feature names shown as they appear in the sets from the experiment. For arousal prediction, it is noted that no cross-modal features were generated, only intra-modal interactions were found on both corpora. This provides some explanation on why feature fusion of eye or eye and face features with the head feature set does not always improve arousal prediction performance from visual input. There may not be as many feature interactions that aid arousal prediction feature representation from the visual modalities. In the case of speech, it is long known that this modality performs well for arousal prediction [35]. This experiment provides further evidence of this in Table 5.2 (a), however, this time an explicit feature candidate for the interactions that could occur inside deep learning algorithms is given. It is interesting to note that there was a head-based intra-modal feature found for arousal on both RECOLA [16] and SEMAINE [11].

For valence, Table 5.2 (b) shows a cross-modal feature candidate generated from face and eye signal interaction on [16]. There were no interaction features found on the SEMAINE [11] corpus. This finding suggests that multimodality is more important for valence than arousal as this was the only case where a cross-modal feature interaction arose. The interaction generated on RECOLA [16] consisted of face, AU12_SD, and eye gaze, delta_gaze_angle_x_quartile_3, features. Specifically, this interaction feature means that measurements of the intensity of raising the lip corners (smiling, while not showing the teeth) scaled by an eye gaze left/right angle measurement, may benefit valence prediction. Eye and smile interactions for some displays of affect have been known to exist for some time, [59], and this result provides a cross-modal feature candidate for such interactions.

The effect of the generated feature interactions when included with the best feature fusion DNN systems from the previous section are given in Table 5.3. Only

Table 5.3: Feature Fusion DNN Continuous Affect Prediction Validation Set Results Where Interaction Features Were Incorporated on (a) RECOLA and (b) SEMAINE with Improvement Above Standard Feature Sets Highlighted (†)

(a) RECOLA

	Arousal		Valence	
Modalities	CCC	*N* Features	CCC	*N* Features
Speech & head	.705	681	-	-
Speech & eye gaze	.720	87	-	-
Speech & face	.726	130	-	-
Head & eye gaze	.566†	550	-	-
Speech & head & eye gaze	.689	472	-	-
Face & head & eye gaze	.508	1,023	.613	1,023
All modalities	**.733†**	**733**	**.656†**	**1,111**

(b) SEMAINE

	Arousal	
Modalities	CCC	*N* Features
Speech & head	**.843†**	**784**
Speech & eye gaze	-	-
Speech & face	-	-
Head & eye gaze	.469	766
Speech & head & eye gaze	.632	469
Face & head & eye gaze	.379	934
All modalities	.620†	1,106

modality combinations where feature interactions were generated were evaluated. The table shows that the interaction measures were largely ineffective for arousal prediction, with performance degradation in 5 out of 7 cases in Table 5.3 (a) and 3 of 5 cases in Table 5.3 (b). Adding the interaction measures improved the fusion of all modalities on RECOLA [16], a relative performance increase of 3.98%, but this CCC (0.733) still fell short of the best-performing feature fusion system for arousal (speech & face, CCC = 0.740). In Table 5.3 (b), it can be observed that there was a noticeable performance in speech & head system performance after the inclusion of the intra-modal head-based interaction feature. This particular system achieved a CCC of 0.843 on the SEMAINE [11] validation set, this was the best feature fusion arousal performance achieved on this corpus, a 19.41% relative CCC improvement compared to the next best feature fusion system. This head interaction feature also improved system performance when added to the multimodal system that incorporated all the available modalities. These results once again show the potential that head-based features have for arousal prediction from visual features. A performance increase for valence prediction was observed with the cross-modal interaction feature as part of the multimodal system that included all modalities, a relative CCC increase of 2.66%. A 2.85% relative performance degradation occurred when incorporating this feature in the face & head & eye gaze system compared to when it was not added.

5.3.3 Screening-based sensitivity analysis

Model fusion outperformed feature fusion in the validation set experimentation thus far. Therefore, based on the improved arousal and valence prediction performance using model fusion, this fusion method is the focus of the experimental results presented and discussed in this and the following sections. The best-performing arousal and valence model fusion systems incorporated all modalities in the fusion framework. Therefore, as part of the screening-based sensitivity analysis for head- and eye-based feature contributions to the models, unimodal model feature group screening was first performed. This was followed by multimodal evaluations for notably beneficial unimodal feature group removals.

5.3.3.1 Head feature groups sensitivity analysis

The validation set results after unimodal head-based feature group removals are given in Figure 5.4. In Figure 5.4 (a) it is shown that removing head rotation resulted in the largest performance drop (12.59% relative CCC performance reduction), suggesting this group to be the most important from the set for arousal prediction on RECOLA [16]. Further, removing either head rotation y (yaw) or head location x was detrimental to resulting models, yielding 5.77% and 4.72% relative CCC performance reductions on this corpus. A performance increase was observed for the removal of head location z (3.15% increase, CCC = 0.590) for the head arousal system on RECOLA [16]. The largest arousal prediction performance decrease on SEMAINE [11], shown in Figure 5.4 (c), was produced by the removal of head location z features, a 21.91% relative CCC reduction. This is the converse of this feature group's behaviour on RECOLA [16] where a performance increase was observed for the removal of this feature group. Similar to RECOLA [16], however, was that removing either head rotation y or head location x provided performance degradation on the SEMAINE [11] validation set. Removing all head location features from the head feature set provided a 0.75% performance increase on the SEMAINE [11] validation set. The performance increases that were observed on both corpora were greater than the average CCC + 1SD for that corpora and were further evaluated in model fusion.

For valence prediction on the RECOLA [16] validation set, large relative CCC performance decreases were observed for the removals of head location (11.73%), head rotation (10.56%), head rotation z (roll) (7.63%) and head rotation y (6.45%) feature groups. These performance changes are shown in Figure 5.4 (b). Increases in valence prediction performance during screening were observed for removals of head location y (4.99% increase, CCC = 0.358), head location x (3.81% increase, CCC = 0.354) and head rotation x (pitch) (1.17% increase, CCC = 0.345) features

on this corpus. Only the first two feature group removals provided performance increases of 1SD or more above the average CCC and were further evaluated in model fusion. On SEMAINE [11], removing all head location or just the head location y features provided the same degradation, 86.51% relative (CCC = 0.039), for valence prediction on the validation set. These results are depicted in Figure 5.4 (d). This suggests that the majority of performance degradation after removing head location features was caused by the head location y group on this corpus. CCC performance decreases were also observed for head rotation (49.83%), head rotation z (roll) (31.83%) and head rotation y (34.60%) feature group removals on SEMAINE [11]. A 25.61% (CCC = 0.363) performance increase was observed on SEMAINE [11] after the removal of the head location z features. This performance increase was greater than the average CCC + 1SD for this affect dimension and was further evaluated in model fusion.

From these results, head rotation appears important for affect prediction using the proposed sets, with head rotation y the most important within that group, across affect dimensions and corpora. An interesting difference across the corpora and affect dimensions was the importance of head location z features. These features appear beneficial (performance degradation when they were removed) for arousal prediction on SEMAINE [11] but not on RECOLA [16]. Alternatively, the head location z features appear important for valence prediction on RECOLA [16] but this is not so for SEMAINE [11]. While there are some similarities that can be observed across the corpora and affect dimensions, this sensitivity analysis indicates that different head features can have different importances in different social situations. Based on this, efforts in prediction domain adaptation, feature selection specifically, can be rewarded by providing suitable feature sub-sets for different social situations.

5.3.3.2 Eye feature groups sensitivity analysis

The unimodal eye-based screening sensitivity analysis results are shown in Figure 5.5. The arousal experimentation shows that all feature groups contributed positively to model performance on RECOLA [16], there were no performance increases observed after removing any feature groups on this corpus. The most important feature groups as evaluated on RECOLA [16], and shown in Figure 5.5 (a), are gaze angle x, the left to right gazing angle, eyes fixated and gaze approach. In the latter two cases of feature group removal, there was only one feature removed as that was all that was selected for these feature groups using mutual information feature selection. However, their removal during screening caused notable arousal prediction performance degradation. These features were gaze fixation time ratio and gaze approach time in seconds mean and when removed, CCC decreases of 25.93% and

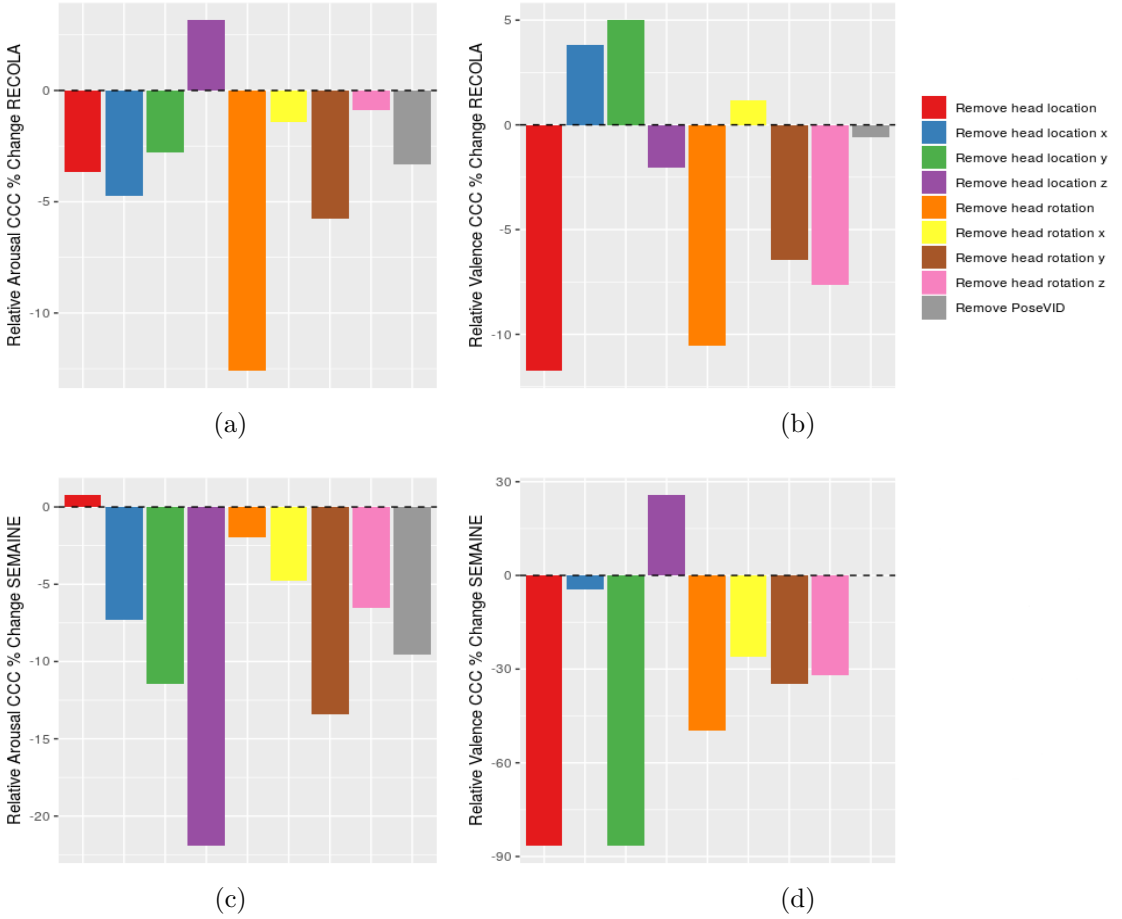


Figure 5.4: Head-based prediction relative validation set CCC % change under different feature group screening conditions on RECOLA for (a) arousal and (b) valence, and on SEMAINE for (c) arousal and (d) valence. The head-based feature sets, prior to screening, provided validation set CCC scores of 0.572 for arousal and 0.341 for valence on RECOLA and 0.676 for arousal and 0.289 for valence on SEMAINE.

26.19% respectively compared to the original feature set were observed. Gaze angle x and eyes fixated were also shown to be important feature groups for validation set prediction on SEMAINE [11] where relative CCC decreases of 35.01% and 28.54% respectively were found after their removal. These results are depicted in Figure 5.5 (c). It can also be seen in this figure that the eye blink and eye closure features were important for arousal prediction from the eye signals on SEMAINE [11]. The removal of the gaze angle y features were the only feature group that caused a performance increase on the SEMAINE [11] validation set. A relative performance increase of 1.44% was observed which improved arousal prediction from the eyes to a CCC of 0.423, this removal was also evaluated at the model fusion stage.

The eye-based valence screening analysis results are shown in Figure 5.5 (b) and (d). It can be seen that the algorithm for model generation benefited by inclusion of gaze angle x, direct gaze and eye blink intensity based feature input on RECOLA

[16]. Of these three feature groups, the direct gaze feature group appears to be particularly important. The CCC performance decrease when this group was removed, 31.58%, was caused by just two features, direct gaze time in seconds max and direct gaze time ratio. Performance increases for valence prediction on RECOLA [16] were observed after the removal of eyes fixated (7.72% increase, CCC = 0.307) and gaze angle y (11.58% increase, CCC = 0.318) feature groups respectively. Only the gaze angle y group removal provided a performance benefit of 1SD or above compared to the average CCC for this group and was further evaluated for model fusion. For the SEMAINE [11] evaluations, all the feature groups appeared to have provided benefits to the learning algorithm. The most important feature groups were gaze angle x, eye gaze approach and eyes fixated feature groups, and their removals provided 116.84%, 107.72% and 94.75% relative CCC performance decreases when removed respectively. For eye gaze approach and eyes fixated features the performance degradation observed was provided by three features and one feature respectively. The eye fixation feature in question is eye fixation time ratio.

5.3.3.3 Model fusion sensitivity analysis

The final results of the screening-based sensitivity analysis, the model fusion evaluations, are detailed in Table 5.4. The valence results in section (b) of this table are given for both individual feature group removals and combined, based on improvement for each individual group removal, feature group removals on RECOLA [16]. For arousal on RECOLA [16], in Table 5.4 (a), it can be seen that removing head location z provided improved model fusion performance, a validation set CCC of 0.850 on the validation set. Removing the gaze angle y features only provided the best valence model fusion performance on RECOLA [16] as can be seen in Table 5.4 (b). This was thought to have occurred due to a reduced intra-training feature correlation by removing these features alone and not the head location x features. This was not confirmed by correlation analysis, however, where the average r across the input training vectors was 0.377 after removing the gaze angle y feature group from its unimodal model; average r was 0.346 when both head location x and gaze angle y were removed. In addition, the average feature-with-target valence correlations were also computed and improved with the removal of both of these feature vectors (removal of gaze angle y feature group $\mu r = 0.407$, removal of both head location x and gaze angle y feature groups $\mu r = 0.414$). To be sure that there were no non-linear effects unaccounted for, average feature-with-feature and feature-with-target MI values were also assessed and the results of the correlation analysis were replicated. Despite what appears to be an improved feature vector when both feature groups were removed, the neural network was able to learn a better representation

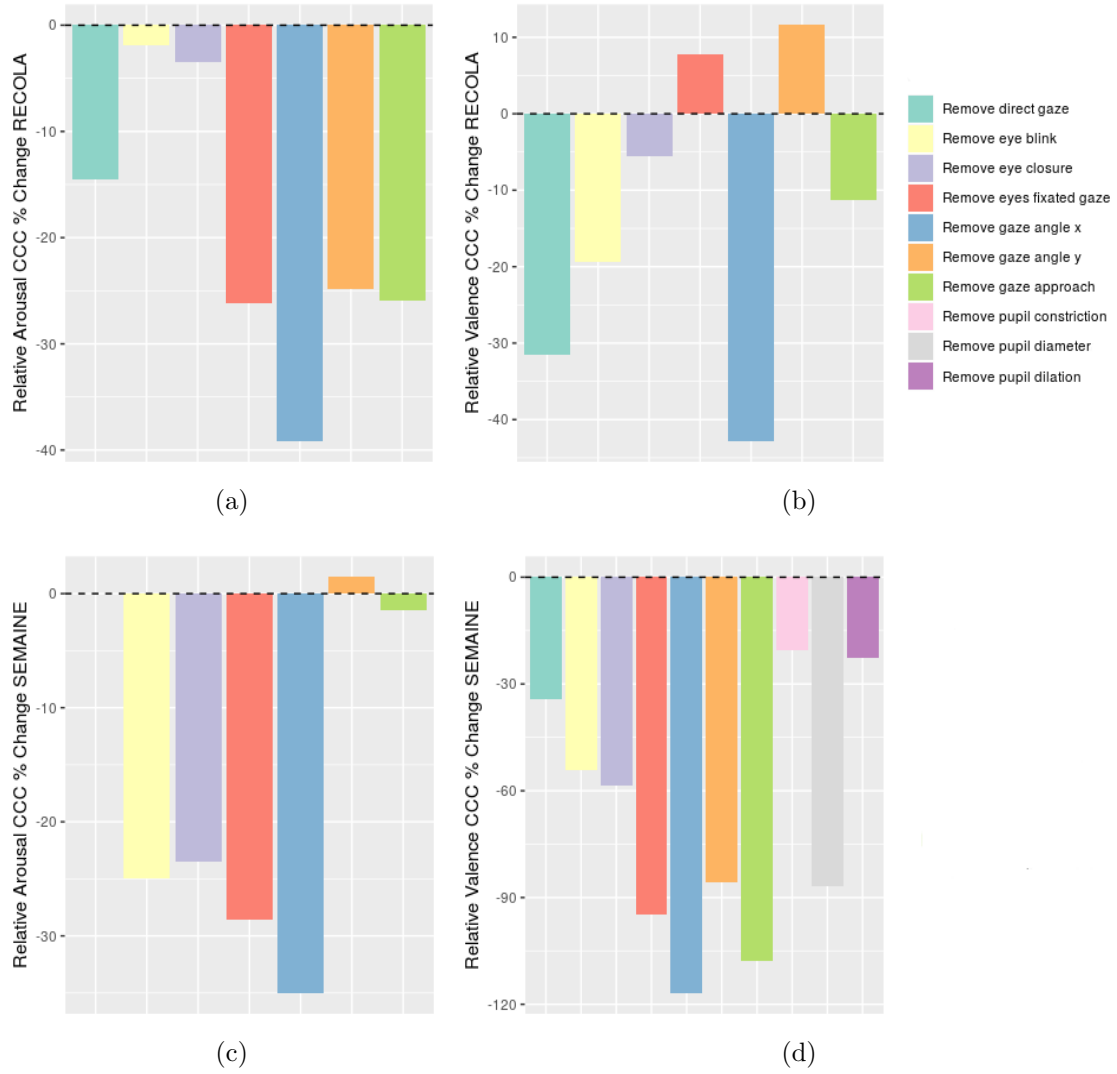


Figure 5.5: Eye-based prediction relative validation set CCC % change under different feature group screening conditions on RECOLA for (a) arousal and (b) valence, and on SEMAINE for (c) arousal and (d) valence. The eye-based feature sets, prior to screening, provided validation set CCC scores of 0.378 for arousal and 0.285 for valence on RECOLA and 0.417 for arousal and 0.285 for valence on SEMAINE.

with only the gaze angle y group removed. Therefore, the removal of gaze angle y feature group only from its unimodal model was chosen for the rest of the valence experiments on RECOLA [16].

The model fusion evaluations on the SEMAINE [11] validation set are shown in Table 5.4 (c) and (d). Only the gaze angle y feature group removal was able to improve the model fusion performance for arousal prediction on SEMAINE [11]. This removal was therefore retained for the rest of the arousal prediction experiments on this corpus.

Table 5.4: Model Fusion DNN Validation Set CCC Results on RECOLA for (a) Arousal and (b) Valence and SEMAINE for (c) Arousal and (d) Valence After Removing Feature Groups Causing Unimodal Model CCC Performance Increases of $\geq$ 1 SD Above the Average CCC for That Modality and Affect Dimension (Δ CCC in the table signifies the relative percentage change in CCC compared to when no feature group removals were applied, N/A)

(a) RECOLA Arousal

Features Removed	CCC	N features	Δ CCC
N/A	.834	1,010	-
Head location z	**.850**	**882**	**2.04%**

(b) RECOLA Valence

Features Removed	CCC	N features	Δ CCC
N/A	.694	998	-
Head location x	.695	870	0.14%
Head location y	.681	870	-1.44%
Gaze angle y	**.723**	**983**	**4.18%**
Head location x and gaze angle y	.721	855	3.89%

(c) SEMAINE Arousal

Features Removed	CCC	N features	Δ CCC
N/A	.859	1,024	-
Head location	.857	640	-0.23%
Gaze angle y	**.862**	**1,008**	**0.35%**

(d) SEMAINE Valence

Features Removed	CCC	N features	Δ CCC
N/A	**.843**	**648**	**-**
Head location z	.820	620	-2.73%

5.3.4 Teacher-forced learning

The validation set results for valence prediction incorporating teacher-forced learning (arousal annotations as an additional training and validation feature) are provided in Table 5.5. Due to delays and short-stopping introduced by the Gaussian smoothing where the window tails were not used, padding was required prior to using arousal predictions as features on the test set. Arousal predictions were therefore set to 0.0 for the multi-stage regression method where 60 predictions were missing at both the start and end of the final test set.

The results of TFL-MSR showed this method to be effective on the validation set for the unimodal models on RECOLA [16] as shown in Table 5.5 (a). The teacher-forced feature always improved unimodal prediction when compared to either standard regression or MTL on this corpus. MTL consistently resulted in a decrease in model performance compared to standard regression. This is thought to have occurred due to the feature selection methods applied, which were aimed at retaining good feature-with-target information for valence, not arousal. The features were therefore not optimal for half of their required prediction task in the MTL approach. In the RECOLA [16] validation set experiments the benefit of using TFL-MSR for valence prediction was not always certain. Model fusion performance was reduced by 0.97% compared to standard regression when the teacher-forced method was used for

all unimodal models for fusion. This is thought to have occurred because each of the unimodal models received nearly identical arousal annotations (only D_s parameters applied differentiated the information) in training and validation, which resulted in more correlated unimodal predictions (less unique information for later fusion) for model fusion input. This suspicion was confirmed using correlation analysis. It was found that the average r across TFL-MSR unimodal model prediction pairs was 0.545 while the average r for the unimodal prediction pairs using the standard method was 0.377. Clearly, while the teacher-forcing method can benefit prediction, care is required so as not to overuse the teacher-provided information. Based on this observation, TFL-MSR was not used for head and face unimodal models for fusion as it did not increase model performance as much as for the speech- and eye-based models. As seen in Table 5.5 (a), these removals were beneficial, providing the best-performing valence model on RECOLA [16], a validation set CCC of 0.740, a 2.35% relative improvement compared to standard valence regression.

The TFL-MSR technique was shown to be ineffective on the SEMAINE [11] corpus. In Table 5.5 (b) it can be observed that TFL-MSR regularly performed worse than the standard regression or MTL approaches to valence prediction on the validation set. This was thought to have occurred due to a low correlation between arousal and valence on the validation set compared to the training set. This suspicion turned out to be true. While this correlation was high on the SEMAINE [11] training set, $r = 0.755$, as explored in Section 3.2.4.3, the validation set correlation between arousal and valence on SEMAINE [11] was only 0.1. This provides an explanation as to why TFL-MSR did not provide a performance benefit on SEMAINE [11]. Even if arousal values were predicted with high fidelity on this corpus and partition, these arousal values simply do not appear indicative of valence on the validation set.

Following the evaluation of the multimodal systems incorporating eGeMAPS [21], similar evaluations were performed using the ComParE [31] LLD-based functionals. This provides a perspective where a feature set with no valence-oriented feature selection was evaluated using TFL-MSR. Due to the previous poor result for TFL-MSR on SEMAINE [11] this method was only evaluated on RECOLA [16]. However, ComParE [31] LLD-based results that did not use the TFL-MSR method on SEMAINE [11] were also gathered to facilitate ComParE [31] LLD-based model fusion input and unimodal comparison of this feature set with eGeMAPS [21]. The results of this experimentation are given in Table 5.6. Arousal results are also provided in this table as they are required for TFL-MSR at the testing stage but have not yet been acquired for the ComParE [31] LLD-based functionals thus far.

While not the focus of this section, it is interesting to note the differences in the unimodal performances presented in Table 5.6 for the ComParE-based set [31] and the previous eGeMAPS [21] results without feature selection (Section 4.4.2). The

Table 5.5: DNN Valence Validation Set Prediction CCC Results on (a) RECOLA and (b) SEMAINE for Unimodal and Model Fusion of All Modalities Using Standard, MTL, and TFL-MSR Learning Approaches

(a) RECOLA

Modality/Fusion	Standard Regression CCC	MTL CCC	TFL-MSR CCC
Speech	.342	.310	.487
Head	.358	.345	.371
Eye gaze	.318	.286	.506
Face	.560	.546	.597
Model fusion	.723	.697	.716
Model fusion (teacher-forced features for speech and eye model input only)	-	-	**.740**

(b) SEMAINE

Modality/Fusion	Standard Regression CCC	MTL CCC	TFL-MSR CCC
Speech	.160	.129	-.018
Head	.289	.328	.269
Eye gaze	.219	-.054	.075
Face	.160	.101	.109
Model fusion	**.843**	.690	.831

unimodal ComParE-based [31] results were always inferior for arousal and valence on both RECOLA [16] and SEMAINE [11]. The unimodal eGeMAPS [21] systems previously obtained CCC scores of 0.741 and 0.314 for arousal and valence respectively on RECOLA [16], and 0.679 and 0.160 for arousal and valence respectively on SEMAINE [11] (Figure 4.11). Of course, while eGeMAPS [21] seems superior for unimodal prediction from these evaluations, this does not guarantee its success as part of a multimodal prediction system. This is the subject of the remaining affect predction evaluations.

Contrary to the previous results for TFL-MSR (Table 5.5), Table 5.6 (a) shows that TFL-MSR caused a large performance decrease when combined with the ComParE-based [31] speech features for unimodal valence prediction. This is thought to have occurred because this feature set without feature selection is naturally suited for arousal prediction, not valence. In this case, providing the teacher-forced arousal feature resulted in an additional feature that was correlated with the other features and did not provide additional information or useful feature interactions for valence prediction. Therefore, the final model fusion result shown in the table does not use the teacher-forced features for the speech-based unimodal model input to the ensemble when using the ComParE-based [31] features. Additionally, teacher-forced features were not used for head and face unimodal model input to model fusion based on the results in Table 5.5. From Table 5.6 (a) it can bee seen that the best performing arousal result thus far was produced using the ComParE-based [31] features as speech input, a CCC score of 0.856.

Table 5.6: Validation Set Results for Systems Incorporating ComParE LLD Functionals Speech: Unimodal, and Model Fusion of All (Best Unimodal) Modalities on (a) RECOLA and (b) SEMAINE (Note: TFL-MSR only employed on RECOLA due to previously poor performance for this method on SEMAINE)

(a) RECOLA

System	Arousal CCC	Valence CCC
Speech	.705	.210
Speech with TFL-MSR	-	.035
Model fusion	.856	.739‡

TFL-MSR used (for eye model input only)‡

(b) SEMAINE

System	Arousal CCC	Valence CCC
Speech	.429	-.001
Model fusion	.843	.831

5.3.5 Final prediction results

The final validation set results are given in Table 5.7 where both speech-based input feature set variations were used on each corpus for this experimental stage. To be noted from the validation set experimentation is that the visual systems never outperformed the multimodal systems that included speech, hence, they are not included in this table. It can be seen that the eGeMAPS-based [21] model outperformed the ComParE [31] LLD functionals-based model on the RECOLA [16] validation set for valence prediction, while the reverse is true for arousal. On SEMAINE [11], the models that incorporated the eGeMAPS [21] features always outperformed the models that included the ComParE [31] LLD functionals-based features on the validation set, however, their performances were comparable. Unbiased estimates of model performance were gathered on the test set for these better performing systems that have their validation set results highlighted in bold in Table 5.7. Biased estimates of model performance, test set passes that were carried out for experimental completeness not based on maximum validation set performance, were gathered for the other systems from Table 5.7. The biased estimates are important to fairly compare with other research or evaluate the TFL-MSR method proposed. However, final model performance claims and model selection in this work are based on observable data (i.e. training and validation data) and unbiased performance estimates, with the final test set pass providing only out-of-sample model scoring.

Of the unbiased model performance estimates given in Table 5.7, the arousal test set CCC scores are among some of the best performances that this author is aware of on each corpus. The valence system evaluated on RECOLA [16] needs improvement, however, where a CCC of 0.463 was achieved. The validation set performance of this system is very different to the test set performance (valence validation CCC = 0.740, valence test CCC = 0.463), which suggests validation set overfitting. To remedy this, regularisation is a possible method for model improvement. The issue of poorer test

Table 5.7: Validation and Test Set Results (including biased test set estimates*) for the Best Performing eGeMAPS and ComParE LLD-based Variations of Multimodal DNN Continuous Affect Prediction Systems From the Experiments on (a) RECOLA and (b) SEMAINE (Note: Estimated Group-of-humans Baseline Test Set CCC Scores: RECOLA Arousal = 0.217, Valence = 0.257; SEMAINE Arousal = 0.398, Valence = 0.500)

(a) RECOLA

	Validation		Test	
System	Arousal CCC	Valence CCC	Arousal CCC	Valence CCC
Model fusion (eGeMAPS speech)‡	.850	**.740**	.771*	**.463**
Model fusion (ComParE speech)‡	**.856**	.739	**.812**	.428*

TFL-MSR used valence learning & prediction‡

(b) SEMAINE

	Validation		Test	
System	Arousal CCC	Valence CCC	Arousal CCC	Valence CCC
Model fusion (eGeMAPS speech)	**.862**	**.843**	**.616**	**.436**
Model fusion (ComParE speech)	.843	.831	.404*	.259*

set results for valence prediction is present on SEMAINE [11] also, as can be seen in Table 5.7 (b). However, valence prediction has proved to be very challenging on this corpus for other authors also [15], [125]. From the unbiased model performance estimates in Table 5.7 it can also be observed that valence prediction on SEMAINE [11] was the only scenario where an automatic system did not exceed the minimum practical performance baseline. The value, for this affect dimension on SEMAINE [11] is an estimated group-of-humans CCC of 0.500 as calculated in Section 3.2.4.1.

To compare the TFL-MSR method to standard regression on the test set, a result for the best performing eGeMAPS [21] valence system that did not use this method was obtained on RECOLA [16]. The result of this was a test set CCC of 0.379. This provides some evidence of the benefit of TFL-MSR compared with the standard approach for valence prediction in some situations. Specifically, when arousal and valence share a correlation relationship on the test set that is somewhat similar to the observed relationship on the training set. Put differently, a lack of train-test mismatch is required for this relationship. The best unbiased estimate of model performance for the TFL-MSR approach reached a CCC of 0.463 on RECOLA [16] compared with a CCC of 0.379 where arousal values/predictions were not used for learning/prediction. This 22.16% relative increase in CCC was provided with only two extra features for learning and prediction as well. Arousal annotations or predictions were only used for speech- and eye-based model input as part of the overall model fusion ensemble.

The test set results obtained, compared to similar approaches from the literature, are provided in Table 5.8. The unbiased estimates of arousal model performance achieved in this work on each corpus, Table 5.8 (b) and (c), compares favourably to the published work used for comparison. Also, the biased, in terms of this experiment, arousal result that used eGeMAPS performed well in comparison with the

other systems that also considered these features as part of multimodal input (Table 5.8 (a)). It is noted that the final feature vectors used in this work are generally larger than those of the referenced authors. A benefit of this work compared to those referenced Table 5.8 (a) and (b), however, is the lack of physiological data input features used, which required sensors to be attached to subjects. Moreover, He et al. [13] and Brady et al. [14] used additional speech feature vectors besides eGeMAPS [21] in their approaches. It is noted that the best unbiased valence prediction performance from this work, Table 5.8 (a), is poor in comparison to the other research work. The results in Table 5.8 (b) corroborate these findings where poorer performance compared to Ringeval et al. [12] was found for valence prediction.

Table 5.8 (b) and (c) contains comparative work from authors that used face and head [12] and face/head pose- and eye-influenced feature [15] input as part of their multimodal continuous affect prediction systems. It is clear from Table 5.8 (b) that the head (PoseVID-adv) and eye (eGazeVID) features developed in this work and included in this system are advantageous when included for multimodal audio-video input for arousal prediction. The valence result is less convincing, however, when comparing against the result of [12] in this table section. The additional performance that they achieved may have been due to the use of physiological data from subjects, as previously mentioned, or due to the more complex LSTM-RNN model used. For the SEMAINE [11] corpus, it can be seen in Table 5.8 (c) that the approach taken in this work performed well when compared with the other published work's unbiased test set results [15]. Unfortunately, this performance came at the technical cost of using more input features and a more complex fusion approach than Han et al. [15]. A disadvantage of their work is the complex combination of both SVR and BLSTM-RNN that they used that ultimately performed quite poor for valence prediction. Model complexity aside, the prediction performance achieved on SEMAINE [11] again shows that the proposed features can positively contribute to continuous affect prediction. PoseVID-adv and GazeVID were used as part of arousal model input, and PoseVID and EyeVID were used as part of valence mode input on this corpus.

It can be seen from Table 5.8 that the DNN-based arousal models of this work were able to match or outperform the LSTM-RNN-based models. This provides some empirical results that recurrence, and thus a more complex model, is not required for continuous arousal prediction. Other authors have also shown non-recurrent models to perform well compared to LSTM-RNN-based models for arousal prediction [129]. Recurrence may still be required for effective valence prediction, however, how much recurrence (100 frames? full sequence?) and alternatives to deep learning recurrent models will be the subject of future study. There is of course a caveat that for the simpler DNN-based models presented in this work, an appropri-

Table 5.8: Multimodal Fusion DNN CCC Results Obtained in This Work on the Test Set Compared Against Related Research That Used: (a) eGeMAPS on RECOLA, (b) ComParE LLD Functionals for Speech (and head-based visual input) on RECOLA and (c) Multimodal Audio-video Approaches on SEMAINE

(a) eGeMAPS speech contribution to multimodal systems on RECOLA

Arousal	Valence	*N* Features	ML Algorithm(s)	Fusion	Authors
.747	.609	102 (arousal), 176 (valence)	BLSTM-RNN	Model fusion	[13]
.770	**.687**	955	SVR, CNN+RNN, LSTM-RNN	Kalman filter	[14]
.753	.399	84 (arousal), 283 (valence)	DNN	Feature fusion	This work
.771	.463	882 (arousal), 985 (valence)	DNN	Model Fusion	This work

(b) ComParE LLD functionals speech and head-based visual input included as contribution to multimodal systems on RECOLA

Arousal	Valence	*N* Features	ML Algorithm(s)	Fusion	Authors
.804	**.528**	1,150 (arousal), 1,420 (valence)	LSTM-RNN	Decision fusion	[12]
.812	.428	1,438 (arousal), 1,562 (valence)	DNN	Model fusion	This work

(c) Multimodal audio-video approaches on SEMAINE

Arousal	Valence	*N* Features	ML Algorithm(s)	Fusion	Authors
.346	.026	75	SVR & BLSTM-RNN	Decision fusion	[15]
.616	**.436**	1,008 (arousal), 648 (valence)	DNN	Model fusion	This work

ate gold standard time-shift for the annotations must be found for model training. Finding high-performing gold standard time-shift values for different modalities that do not overfit to specific situations (e.g. the training set) may be problematic. Researchers have tried to learn delays automatically as part of prediction models but large differences in test set results compared to validation set results were found [128]. More research is therefore needed for automatic learning of these delays. A final observation on the results in Table 5.8 is that the currently developed models from the literature and this work do not lend themselves well to model interpretability. This is due to the generally large amount of nonlinear transformations and feature interactions inside the neural networks used [162]. It is believed, however, that the head- and eye-based features of this work can be deployed successfully with other more interpretable models if required. The use of interpretable models is important for future high-stakes affective computing systems, such as those that might be applied in healthcare [163], for example.

5.3.6 Feature retrospective

Prior to looking back at the head- and eye-based features' relationships with arousal and valence, relationship scores from the speech and face modalities were obtained for comparison. To achieve this, the best linear and nonlinear feature-with-target relationships from these modalities that contributed to the final systems were gathered

Table 5.9: The Highest-performing of the Speech- and Face-based Feature Rankings by Correlation (absolute value) and Mutual Information (nats) on RECOLA for (a) Arousal and (b) Valence, and SEMAINE for (c) Arousal and (d) Valence, Calculated on the Corpora Training Sets

(a) RECOLA Arousal

Feature	r	Feature	MI
pulse-code modulation RMS energy simple moving average derivative SD	.756	pulse-code modulation RMS energy simple moving average derivative min	.737

(b) RECOLA Valence

Feature	r	Feature	MI
action unit 12 intensity max	.636	action unit 12 intensity max	.708

(c) SEMAINE Arousal

Feature	r	Feature	MI
spectral flux UV simple moving average (applied on 3 non-zero frames only) arithmetic mean	.533	action unit 12 intensity max	1.184

(d) SEMAINE Valence

Feature	r	Feature	MI
action unit 7 intensity mean	.609	action unit 17 intensity max	1.149

on the corpora training sets. The results of these calculations are given in Table 5.9. It can be seen in the table that features based on the first-order derivative of the digitised speech signal energy share the highest feature-with-arousal relationships on RECOLA [16]. For valence, the same feature, based on the AU12 intensity, an action unit based on lip corner puller (author's interpretation: smile with no teeth), shares the highest relationships with valence on this corpus. On SEMAINE [11], shown in Table 5.9 (c), a spectral flux acoustic signal feature provided the best correlation with arousal and AU12 intensity max shared the largest MI or nonlinear relationship with this affect dimension. Features based on AU7 and AU17 shared the highest linear and nonlinear relationships respectively from the speech and face modalities with valence on SEMAINE [11]. AU7 is formally called lid tightener which is interpreted by this author as a drawing together of both sets of eyelids that does not result in eye closure. AU17, formally called chin raiser, is a movement of the chin facial muscles that causes a raising of the lower lip centre and slight lip corner depression according to this author.

The top features, as ranked by feature-with-target relationships, from the head modality that were selected for the final continuous affect prediction systems are given in Table 5.10. From this table, the wavelet coefficient-based features appear to be promising features arousal prediction on both corpora. On RECOLA [16], the 3^{rd} level decomposition wavelet coefficients-based ($[1.5625, 3.125]$Hz) features make up the top 5 feature-with-target r relationships, all based on the head location x signal. For the nonlinear relationships with arousal on this corpus, the summaries of simpler head measurements fared well, 7 of the 10 top-performers as ranked by

MI are based on these measures. Many of these top-performers were based on head location y measurements. Scale coefficients from the wavelet decomposition regularly appeared in the arousal MI rankings on RECOLA [16] as well. Scale coeffs l1 head location y max (from $[0, 6.25]$Hz band) ranked 2^{nd} and scale coeffs l2 head location y max (from $[0, 3.125]$Hz band) ranked 4^{th} in terms of MI. On SEMAINE [11], the 4^{th} level decomposition wavelet coefficients-based features ($[3.125, 6.25]$Hz band) comprise 9 out of 10 of the top feature-with-target r ratings and the majority of these relationships arose from decomposition of head location z. The SEMAINE [11] arousal features, Table 5.10 (c), are the only head features that have arousal MI relationships higher than their speech and face counterparts. All the top 5 MI relationships in this table are larger than the relationship calculated for face MI with arousal. Similar to RECOLA [16], the SEMAINE [11] MI relationships were largely provided by the summaries of simpler measurements/calculations from the head signal while scale coeffificent-based features are present in the top-ranked features as well. Different from RECOLA [16], however, is that the z axis location measurement appears more important in terms of MI with arousal on SEMAINE [11].

The head feature relationships with valence are given in Table 5.10, sections (b) and (d). There are some similarities across the arousal and valence relationships on RECOLA [16]. The 3^{rd} level decomposition wavelet coefficient head-based features once more occur regularly in the top 10 features as ranked by r. For the nonlinear feature relationships with valence on RECOLA [16], 6 of the top 10 features as ranked by MI are from the head location z signal. The highest feature relationship with valence, however, is Δ head location y min, a MI of 0.616. This estimated MI is 12.99% less (relative) than the strongest face modality MI with valence. On SEMAINE [11], head location z features make up the majority of top-ranked valence relationships. As was the case for the head arousal features on SEMAINE [11], all the top 5 MI relationships in this table are larger than the relationship calculated for face MI with valence. Surprisingly, head rotation x (pitch) features that were expected to regularly score highly for this affect, based on [26], [65], are not prevalent in the top-ranked feature lists in this analysis.

In summary, these results mean that both left-to-right and up-down head measurements can play important complimentary roles in affect prediction in some situations, for example RECOLA [16], with one contributing more to modelling the linear part, and the other, the nonlinear aspect of arousal from head pose/motion. In other social scenarios, SEMAINE [11] being the example from this work, the pose distance or z axis location measurements can carry more of the affective signal from the head. The results provide some empirical evidence for Schinerla's approach-withdrawal hypothesis [164], [165], that one would approach or withdraw

Table 5.10: Top Head-based Features Selected for the Final Model Fusion DNN System Ranked by Correlation (absolute value) and Mutual Information on RECOLA for (a) Arousal and (b) Valence, and SEMAINE for (c) Arousal and (d) Valence, Calculated on the Corpora Training Sets

(a) RECOLA Arousal

Feature	r	Feature	MI
wavelet coeffs l3 head location x IQR 1-3	.420	head location y max	.599
wavelet coeffs l3 head location x quartile 3	.410	scale coeffs l1 head location y max	.581
wavelet coeffs l3 head location x IQR 2-3	.404	Δ head rotation y max	.577
wavelet coeffs l3 head location x quartile 1	-.400	scale coeffs l2 head location y max	.561
wavelet coeffs l3 head location x IQR 1-2	.401	head location y quartile 3	.553
wavelet coeffs l4 head location x IQR 1-3	.385	Δ head location x max	.551
Δ head location x IQR 1-3	.384	Δ head rotation x max	.549
wavelet coeffs l4 head location x quartile 1	-.380	Δ head rotation y min	.545
Δ head location x quartile 3	.372	scale coeffs l1 head location y quartile 3	.543
wavelet coeffs l4 head location x IQR 1-2	.368	head rotation y min	.541

(b) RECOLA Valence

Feature	r	Feature	MI
Δ head location y IQR 1-2	.363	Δ head location y min	.616
Δ head location y IQR 1-3	.349	head location z max	.607
wavelet coeffs l3 head location x IQR 1-3	.336	Δ head rotation x min	.606
wavelet coeffs l3 head location y IQR 1-3	.332	Δ head location z min	.602
Δ head location x quartile 3	.331	Δ head location z IQR 1-3	.595
wavelet coeffs l3 head rotation x quartile 3	.327	Δ head location x max	.593
wavelet coeffs l3 head location x IQR 2-3	.323	head location z quartile 3	.591
wavelet coeffs l3 head rotation y IQR 1-2	.323	scale coeffs l1 head location z quartile 3	.589
wavelet coeffs l3 head location x quartile 1	-.322	head location z min	.588
wavelet coeffs l3 head location y quartile 3	.320	Δ head rotation y max	.587

(c) SEMAINE Arousal

Feature	r	Feature	MI
wavelet coeffs l4 head location z IQR 1-3	.496	head location z max	1.239
wavelet coeffs l2 head location z ZCR	.480	head location z min	1.226
wavelet coeffs l4 head location z quartile 1	-.480	Δ head location z min	1.198
wavelet coeffs l4 head location z IQR 1-2	.471	head location y max	1.197
wavelet coeffs l4 head location z quartile 3	.466	head location z quartile 3	1.186
wavelet coeffs l4 head location z IQR 2-3	.444	scale coeffs l1 head location z quartile 3	1.183
wavelet coeffs l4 head location x IQR 1-3	.435	scale coeffs l2 head location z quartile 3	1.178
wavelet coeffs l4 head location x quartile 3	.425	scale coeffs l1 head location z max	1.167
wavelet coeffs l4 head location y IQR 1-3	.421	scale coeffs l3head location z quartile 3	1.167
wavelet coeffs l4 head rotation y IQR 1-3	.420	head location z median	1.165

(d) SEMAINE Valence

Feature	r	Feature	MI
head rotation x quartile 3	.492	head location z max	1.251
Δ head rotation x IQR 2-3	-.473	head location z min	1.207
head rotation x median	.470	head location z median	1.155
head location y mean	-.467	head location z quartile 1	1.152
head rotation x mean	.467	head location z quartile 3	1.152
head location y median	-.465	Δ head location z IQR 1-3	1.135
head location y quartile 1	-.456	head location y max	1.127
head location y quartile 3	-.448	Δ head location z max	1.127
head location y min	-.433	head rotation x max	1.103
head rotation x quartile 1	.429	head location z mean	1.102

from stimuli depending on its content and intensity. In the observed relationships from the experimentation, the approach or withdrawal of individuals (z axis location measurements) was a reasonable nonlinear signal of the valence observed by the annotators across both corpora.

In Table 5.11, the top 10 eye features as ranked by feature-with-target correlation and MI on each corpus are given. These results show that the eye-based features

generally share a better relationship with arousal than valence on RECOLA [16] while the reverse is true for SEMAINE [11]. Gaze angle measurements are clearly important for arousal prediction from the eyes across both corpora according to Table 5.11. These features make up 8 of the top r values and 9 of the top MI measurements from a possible 10 on RECOLA [16]. On SEMAINE [11], 8 of the top 10 features as rated by r and 8 of the top 10 as rated by MI are gaze angle-based. In comparing the relationships across the corpora, gaze angle y measurements were highly rated more often on SEMAINE [11] whereas the relationships on RECOLA [16] are more mixed between x and y gazing angles.

The gaze angle-based features were less dominant in relation to the top valence relationships. There were 5 of 10 appearances for these feature groups in Table 5.11 (b) for both r and MI measurements on RECOLA [16] respectively. The same proportions for the gaze angle feature group's relationships compared to all other relationships were found on SEMAINE [11] also, and are shown in Table 5.11 (d). Interestingly, the top 4 eye features as rated by MI on SEMAINE [11] were all better that the top-rated face feature for this corpus and affect. In spite of the similarities across the corpora, the other descriptors that are highly rated for valence include features related to specific forms of gaze (eyes closed, eye blink intensity, direct gaze, gaze approach) on RECOLA [16] and pupil-based measures on SEMAINE [11]. It is possible that the differences in the observed relationships for valence across the corpora can be attributed to the professional lighting and recording equipment used in SEMAINE [11]. This is a drawback for consideration of deploying the proposed pupil features in everyday situations. Lighting and recording conditions, unless highly controlled, appear to render the pupil features unusable.

5.3.7 General discussion

In the experiments, the head- and/or eye-based features were able to improve upon the best unimodal performers of Chapter 4 when model fusion was used. Surprisingly, feature fusion, when performed early, was generally unable to improve upon the best unimodal systems of Chapter 4. The best systems from the fusion experimentation were always the model fusion systems that used all modalities. These systems, or the other multimodal systems that included head & eye input, always showed clear performance improvements above the next best performing bimodal model fusion system. The experiments therefore show the efficacy of the proposed feature sets as auxiliary features for multimodal continuous affect prediction in audio-video when fused appropriately. Different modalities have different response times in humans [166], [167], which perhaps led to model fusion's success where modalities are integrated with arousal or valence at different times for learn-

Table 5.11: Top Eye-based Features Selected for the Final Model Fusion DNN System Ranked by Correlation (absolute value) and Mutual Information on RECOLA for (a) Arousal and (b) Valence, and SEMAINE for (c) Arousal and (d) Valence, Calculated on the Corpora Training Sets

(a) RECOLA Arousal

Feature	r	Feature	MI
gaze angle x quartile 3	.324	gaze angle x max	.551
gaze angle x max	.319	Δ gaze angle y min	.548
gaze angle x median	.288	Δ gaze angle x max	.532
gaze angle x mean	.286	Δ gaze angle y max	.529
gaze angle y min	-.265	Δ gaze angle x min	.522
direct gaze time ratio	.260	gaze angle y min	.501
direct gaze time secs total	.245	gaze angle x min	.488
gaze angle x SD	.244	eye blink intensity max	.479
Δ gaze angle y IQR 1-3	.236	gaze angle y max	.458
Δ gaze angle y quartile 3	.232	gaze angle x quartile 3	.427

(b) RECOLA Valence

Feature	r	Feature	MI
gaze angle x quartile 3	.276	gaze angle x max	.548
gaze angle x max	.275	eye blink intensity max	.486
gaze angle x quartile 1	.217	gaze angle x quartile 3	.395
direct gaze time ratio	.192	gaze angle x quartile 1	.385
direct gaze time secs max	.188	Δ gaze angle x SD	.312
gaze angle x SD	.168	eyes closed time secs min	.306
eye blink intensity IQR 2-3	.166	direct gaze time secs max	.302
eye blink intensity max	.152	gaze angle x SD	.273
gaze angle x IQR 1-2	.104	eyes closed time ratio	.266
eyes closed time secs min	-.100	direct gaze time ratio	.236

(a) SEMAINE Arousal

Feature	r	Feature	MI
gaze angle y SD	.309	gaze angle y max	1.157
delta gaze angle y IQR 2-3	-.280	Δ gaze angle y max	1.136
delta gaze angle y IQR 1-2	-.276	Δ gaze angle x min	1.121
gaze angle y IQR 2-3	.252	gaze angle x min	1.118
gaze angle y max	.251	eyes closed time secs max	1.002
gaze angle y kurtosis	-.249	gaze angle y mean	.863
gaze angle x min	-.238	gaze angle x quartile 1	.794
gaze approach time secs mean	.235	eyes closed time secs min	.660
gaze approach time secs max	.221	gaze angle y SD	.611
gaze angle y mean	.212	Δ gaze angle y skewness	.591

(b) SEMAINE Valence

Feature	r	Feature	MI
Δ gaze angle y IQR 1-3	-.592	gaze angle y min	1.184
Δ gaze angle y quartile 3	-.591	gaze angle y max	1.169
Δ gaze angle y IQR 2-3	-.589	eye blink intensity max	1.157
Δ gaze angle y quartile 1	.583	gaze angle x max	1.153
Δ gaze angle y IQR 1-2	-.577	delta pupil diameter mm min	1.147
pupil dilation time secs mean	.534	Δ gaze angle y min	1.127
wavelet coefficients l1 IQR 2-3	-.527	Δ pupil diameter mm max	1.126
wavelet coefficients l1 quartile 3	-.526	pupil diameter mm min	1.125
wavelet coefficients l1 IQR 1-3	.521	pupil diameter mm max	1.121
wavelet coefficients l1 quartile 1	-.517	Δ gaze angle x max	1.104

ing. This also highlights future opportunities for early feature fusion. Different modalities could be integrated with arousal or valence at different times, shifting features forward at different rates, modality-wise, prior to training. This could facilitate better feature representations within neural networks for this fusion method.

To search the speech, face, head, and eye multimodal feature space further, cross-modal interaction features were investigated using autoML. Cross-modal features

were not found for arousal in the experiments on both corpora, only within-modality feature interactions were found for this affect dimension. A large performance increase was observed when an intra-modal head-based interaction feature was used for speech & head arousal prediction SEMAINE [11]. This likely occurred due to the head modality's efficacy as a visual predictor of arousal combined with this feature's favourable interactions with the rest of the feature vector inside the neural network. One cross-modal face and eye gaze interaction was found for valence on RECOLA [16]. A performance increase was also observed when using this cross-modal feature for feature fusion-based multimodal valence prediction. This is reasonable because eye signals are often considered to be part of subject's facial displays and annotators most likely experience and perceive some of these shared signals simultaneously. These results showed some promise for exploring interaction features explicitly, in agreement with other literature on multi- and cross-modal affect signalling [24], [59].

Algorithm sensitivity analyses, performed by way of feature group screening, showed head rotation-based features (head rotation y-based features in particular) to be important for both unimodal arousal and valence prediction tasks across both experimental corpora. These features did not appear as often as highly ranked with arousal or valence compared to their head location counterparts during a feature retrospective performed after modelling, however (Table 5.10). This indicates the importance of features with smaller relationships with target values to the feature set as a whole because they contribute information that is useful for the overall modality feature representations. An interesting difference between RECOLA [16] and SEMAINE [11] during the sensitivity analysis was the large difference in performance when head location z features were removed. Removing these features appeared to benefit arousal model performance on RECOLA [16], while a large performance penalty was paid for their removal on SEMAINE [11]. Further, this relationship was reversed for the valence sensitivity analysis where head location z features degraded prediction performance on RECOLA [16] but benefited prediction on SEMAINE [11]. The removal of head location z features also benefited the final multimodal arousal model performance on RECOLA [16]. For the eyes, the sensitivity analysis showed that valence models performed notably better with the exclusion of gaze angle y features for unimodal prediction on RECOLA [16]. A similar relationship was observed on SEMAINE [11], but for the arousal dimension. Furthermore, this removal of gaze angle y-based features from the unimodal models had a positive effect for the multimodal model fusion prediction ensembles. It appears that different head- and eye-based features can have different importances in different social situations for affect prediction. This difference could perhaps be leveraged toward socio-affective context processing in the future as differences in context have been discussed as important for emotion signal disambiguation [45].

The final test set results for arousal prediction compared favourably with other published work for audio-video multimodal affect prediction on both corpora [12]–[15]. Additionally, these results were provided with the advantage of a simpler core learning algorithm. A drawback of the approach taken in this work was the larger input feature vectors compared to the other authors, however. The final valence systems require improvement also, even in light of improved performance compared to an independent multimodal approach [15] on SEMAINE [11]. A different algorithm or more regularisation and/or other hyperparameter alteration for the current DNN algorithm may be required to improve upon the obtained valence results. An interesting finding on the validation valence performances was that the proposed TFL-MSR improved prediction performance on RECOLA [16] but not on SEMAINE [11], due to a low arousal-valence correlation on the latter corpus. The final valence test set result on RECOLA [16] improved by a relative performance increase of 22.16%, compared to when this method was not applied. The experiments also showed that overuse of teacher-forced features is not beneficial for model fusion input and using a valence-oriented feature selection method may be required for the method to work. All the model performances on the test set, except valence on SEMAINE [11], surpassed the group-of-humans CCC performance estimates, therefore showing a practical benefit of the features proposed and models used, in general. Overall, the results highlight the importance of using multimodal approaches in human expression analysis, which has been identified previously [24], [25].

5.4 Conclusion

In this chapter, the proposed head- and eye-based feature sets were evaluated in multimodal settings to answer the research question:

How much of an improvement can head- and eye-based features provide when included in multimodal continuous affect prediction systems?

From the experiments that were carried out it was shown that combining all the available modalities by way of model fusion always performed best for affect prediction. In comparing these multimodal systems against the best of unimodal or bimodal speech and/or face feature input systems, the multimodal systems were 4.91% better for arousal and 18.23% better for valence on RECOLA [16] and 13.18% better for arousal and 74.17% better for valence on SEMAINE [11], relative CCC. These CCC results are given in Table 5.1. It is therefore concluded from these experiments that incorporating head and eye cues as part of multimodal input for affect

prediction can improve arousal performance by 9.05% and valence performance by 46.02% on average.

During further experimentation, automatically generated cross-modal interaction features were not found for arousal on either corpus but one was found for valence on RECOLA [16]. A technique to exploit human arousal annotation knowledge gained by models during training, TFL-MSR, was investigated for affect prediction for the first time. This method was shown to be effective on RECOLA [16] but not on SEMAINE [11]. This technique did not work on SEMAINE [11] due to a very small relationship observed between arousal and valence on the validation set of this corpus. When the arousal-valence relationship is not very small, TFL-MSR allows annotation knowledge to be further exploited on unseen test data using multi-stage regression as evident by the final RECOLA [16] test set results. This provides new opportunities for researchers to take further advantage of arousal annotations provided by subjects to improve valence prediction when the testing conditions are appropriate. The systems developed in this work are based on less complex feed-forward DNN models compared to commonly used LSTM-RNN-based models. The performances obtained for arousal prediction in the experiments compared favourably with other published work and this indicates that recurrence is not a requirement for prediction of this affect dimension. An identified limitation with this work and that of the other published studies was the high degree of complexity of the models used for prediction. This could be problematic should interpretable affect prediction models ever be required. The features developed in this work could be employed in future interpretable continuous affect prediction systems, however, in domains where this may be required, such as healthcare [163].

The models from this chapter provide the final, developed arousal and valence models from this work based on the validation set performances achieved. In the next chapter, a summary of the work presented in this dissertation along with final conclusions and suggestions for future work is provided.

Chapter 6

Summary and Future Work

6.1 Summary

The focus of the research in this dissertation was the improvement of continuous affect prediction using audio and video. This involved developing head- and eye-based features and methodologies for this task. The review of related work, carried out in Chapter 2, showed that head and eye cues were underexplored for continuous affect prediction despite the benefits they can provide [4], [22]–[26], [32], [59], [65], [71], [80], [81]. The speech modality was often used as an upper-end baseline to improve upon in the experiment chapters, Chapters 4 and 5. An estimated affect prediction performance from a group-of-human annotators was also used as a minimum performance benchmark in these chapters.

Chapter 4 involved the proposal for feature vectors from the head, PoseVID and PoseVID-adv, and eyes, GazeVID, eGazeVID and EyeVID, for continuous affect prediction. Different temporal windows for feature extraction, annotator delay compensation for gold standard values and feature selection methods were evaluated using a DNN on the RECOLA [16] and SEMAINE [11] affective corpora to appraise the feature sets. The proposed features were only considered as unimodal input to the DNN and their performances were compared against unimodal speech and face input. In general, an 8-second temporal window for feature extraction performed best in the experiments. The best-performing systems from the experiment were speech, without feature selection, for arousal prediction, a CCC of 0.741, and face, using simple MI-based filter feature selection, for valence, a CCC of 0.560, on RECOLA [16]. On SEMAINE [11], the speech system that employed mRMR feature selection performed best for arousal prediction, achieving a CCC of 0.736, and head input without feature selection, PoseVID specifically, performed best for valence prediction, a CCC of 0.289. The proposed head feature sets performed well for arousal prediction, with PoseVID-adv without feature selection specifically per-

forming second-best for arousal prediction on both corpora, while also exceeding estimated human performance baselines. Another notable performance for arousal overall was that of the GazeVID feature set; it surpassed the human performance baseline on both corpora. Even though PoseVID performed best for valence prediction on SEMAINE [11], it did not surpass human performance. It was therefore concluded that head-based features are suitable for arousal prediction on their own and performance within -15.82% relative CCC compared with speech can be obtained.

In Chapter 5, further evaluation of the proposed feature sets was carried out by way of multimodal fusion, cross-modal feature interaction and TFL-MSR experimentation. The features were compared to and combined with speech and facial features for continuous affect prediction using a DNN on the RECOLA [16] and SEMAINE [11] corpora to achieve this. Model fusion of all modalities performed best from the fusion methods evaluated and performance improvements were observed when head and eye features were included in multimodal systems. On RECOLA [16], model fusion of all modalities provided relative CCC performance increases of 12.57% for arousal prediction and 18.23% for valence prediction compared with model fusion speech & face. On SEMAINE [11], model fusion of all modalities outperformed model fusion speech & face by 13.18% for arousal prediction and 74.17% for valence prediction, all relative CCC. Only one cross-modal interaction feature was found, a face (smile with no teeth) and gaze (left-to-right angle) interaction for valence prediction on RECOLA [16], and this feature was able to improve feature fusion valence performance. A screening-based learning algorithm sensitivity analysis was performed that showed interesting similarities and differences across the affective corpora, and therefore, social situations. Head rotation y(yaw)-based features appeared important across both corpora for arousal prediction while head location z was a poor feature group for arousal prediction on RECOLA [16] but not on SEMAINE. Another interesting difference across the corpora was that removing gaze angle y features benefited valence prediction on RECOLA [16] and arousal prediction on SEMAINE [11], they were a poor feature group for these cases. The TFL-MSR method that incorporated arousal annotations and predictions as a feature was shown to benefit continuous valence prediction on RECOLA [16] but not SEMAINE [11]. TFL-MSR assisted in providing the best-performing valence system on RECOLA [16] but degraded performance on SEMAINE [11], where only a small correlation between arousal and valence was present during evaluation. In this chapter, it was concluded that the proposed feature sets can benefit continuous affect prediction and that TFL-MSR can improve valence prediction performance with appropriate (i.e. the arousal and valence correlation is not small/weak) testing conditions.

6.2 Primary contributions

The two major contributions of the work presented in this dissertation are as follows:

- Feature vectors are proposed from the underexplored head- and eye-based information sources for the purpose of continuous affect prediction. Parameters for use with the proposed features are given and the usefulness of the head-based features was shown for unimodal continuous arousal prediction using DNN. Different performances were obtained for the differing head- and eye-based feature sets of varying sizes and complexities. This shows that the more complex feature sets are not always better; especially notable was the most complex eye feature set, EyeVID, which generally performed worse than its simpler counterparts.
- The head and eye feature sets were combined with speech and facial features for multimodal experiments and cross-modal feature interaction, and a new method for valence learning, TFL-MSR, was investigated. Improved multimodal affect prediction performance for DNN models was observed when the head and eye features were included as part of system input, demonstrating their value as auxiliary information sources from video. One cross-modal interaction feature was found involving face and eye gaze for valence. The TFL-MSR method was shown to improve valence prediction performance when there is not a small correlation relationship present between arousal and valence.

6.3 Final conclusions and perspectives

In Section 1.1.2 the following research question was posed:

For audiovisual communication, how much of an improvement in the continuous prediction of core affect can be achieved by processing the combined cues gathered from an individual's speech, head and eyes?

The multimodal experiment provided results that showed an increase for arousal prediction above the next-best system that incorporated speech ranging from 4.91 to 18.23% relative CCC. By incorporating the proposed features, valence prediction was improved by 18.23 to 74.17% relative CCC. Averaging these experimental results, relative performance improvements of 9.05% for arousal and 46.20% for valence can be obtained by combining the head and eye feature sets with speech, thus providing answers to the research question for each core affect dimension.

From the experiments, it can also be concluded that a DNN can outperform a LSTM-RNN for continuous arousal prediction. Therefore, a simpler model can perform well for this task, albeit with additional processing required in the form of gold standard backward time-shifting. The temporal context modelling capabilities of the LSTM-RNN are therefore questionable, outside of gold standard alignment, for arousal prediction. It may be that the context window of 8 seconds used in this work might be enough to capture the temporally salient features in an affect display as judged by annotators. Five feature sets from head- and eye-based modalities have been proposed in this work, of which four have been shown to generally benefit multimodal continuous affect prediction. The EyeVID feature set showed limited effectiveness, only providing a benefit for valence prediction in the professional recording setting of SEMAINE [11]. A novel method has been proposed for valence modelling, TFL-MSR, and this was shown to provide benefits above standard regression learning for the prediction of valence. The final multimodal models, based on the validation set results achieved in this work, are the Chapter 5 models. Namely, these are the model fusion multimodal model incorporating ComParE for arousal, the model fusion multimodal model incorporating eGeMAPS using TFL-MSR for valence on RECOLA [16] and the eGeMAPS model fusion multimodal models on SEMAINE [11].

The head and eye feature sets generated and evaluated in this work have the potential to assist in affective computing interpretability efforts. This means that with the appropriate model, why predictions are made, based on observed input values, can be understood appropriately skilled individuals. The reason why this can be done is that named entities are provided for the features, guided from previous psychological and affective computing research, that themselves have potential to be interpreted. This stands in contrast to unnamed, nonlinear and (many) feature/activation interaction-dependent automatically-learned neural network activation features. These deep-learned features can perform well for affect prediction and require minimal feature engineering effort [33], [34], however, such features are not interpretable at this time. Interpretable affective computing may be very important in the future, for high-stakes decisions such as those required in healthcare [163], for example.

Some wider implications for the head and eye feature sets researched in this work include their potential applicability for audio-visual pathology research and other speech related tasks. For example, 3 degrees-of-freedom head movement and eye closure and gaze features have been used with speech for multimodal depression classification [5]. The features developed in this dissertation, based on 6 degrees-of-freedom head movement and some eye features not used in [5] may benefit multimodal depression recognition. Head- and eye-based features have also been used

as part of a mixed closed/open-source post-traumatic stress disorder estimation system [4], highlighting another pathology application where the features developed in this research could be applied. It has also been shown that visual features increase speech perception [168] and non-linguistic vocal outburst recognition [169] in the presence of audibile noise, while eye gaze features appear important for automatic conversation analysis [170], [171]. Outside of core affect prediction, the feature sets produced in this work can provide new visual descriptors to study other aspects of audiovisual social interactions and their dynamics.

6.4 Limitations and future work

Some limitations of the work in this dissertation include the small subset of SEMAINE [11] used and the perhaps sub-optimal temporal delays applied to gold standard values used for feature fusion in the final experiment. Based on this, future work could investigate different methods for finding suitable delays for gold standard annotations, including modality-wise asynchronous feature fusion delays. A brute-forcing method of using many easily trained models such as L1 regularised regression [172] could be carried out. This could supply a multiple-modality-time-integration regression model or modality delay hypotheses for other, more complex models. This work could be done on RECOLA [16] and an expanded SEMAINE [11] subset. Further evaluation of the features proposed in this work as feature candidates for interpretable affect prediction models [162] could also be carried out. Also, the implementation of automatic direct gaze detection methods [173] for gathering the direct gaze-based features of this work is another possible future work direction. Finally, the features developed in this work could be compared against other head and eye sets that have been applied to social role [171] and engagement [170] detection in automatic conversation analysis. It is hoped that this research informs and inspires the future development of head- and eye-based continuous affect prediction and other affective and social computing applications.

Appendix A

Software Tools/Packages and Revisions

The R programming language and interpreter [174] was the primary programming language and execution environment for the feature generation carried out in this work. Further software tools, revision numbers and notes for when they were used are provided in Table A.1 along with information on R packages used in the experiments.

Table A.1: Experimental Software Tools/Packages and Revisions

Software	Version	Notes
R	3.4.4	Programming language & interpreter, mean, median, quartile, SD, LR and Pearson's r calculations using R base package
RStudio	1.1.383	Integrated development environment
openSMILE	2.3.0	Speech feature extraction
CURRENNT	0.2 rc1	DNN training and prediction
mplayer	1.1-4.8	mp4 to Waveform Audio File Format conversion for openSMILE
R package, DescTools	0.99.21	CCC analysis
R package, infotheo	1.2.0	MI estimation
R package, e1071	1.6.8	Skewness and kurtosis calculations
R package, ncdf4	1.16	netcdf file writing for CURRENNT
R package, mRMRe	2.0.9	mRMR feature selection
R package, wavelets	0.3.0.1	Discrete wavelet transform
R package, seewave	2.1.0	ZCR and RMS calculations
R package, smoother	1.1	Gaussian smoothing

Acronyms

AU action unit. 40, 41, 65, 81, 82, 103, 117

autoML automatic machine learning. xii, 96, 103, 121

AVEC audio-visual emotion challenge. 22, 37–39, 42, 47, 48, 51, 55

BLSTM-RNN bidirectional long short-term memory recurrent neural network. 33, 39, 42, 44, 46, 48, 50, 52, 59, 115, 116

BoW Bag-of-words. 39, 41, 49

BPTT backpropagation through time. vii, 29, 30

CCC concordance correlation coefficient. i, viii–xiii, 24, 34–39, 41, 42, 46, 49, 50, 57, 58, 63, 81, 83–90, 95, 96, 98, 100–102, 104–114, 116, 123, 125–127, 130

CNN convolutional neural network. 39, 41, 42, 45, 50, 52, 116

ComParE Interspeech Computational Paralinguistics ChallengE. xiii, 37, 38, 93, 94, 99, 111–114, 116, 128

CRM cooperative regression model. 46

CURRENNT Cuda recurrent neural network toolkit. 61, 62, 130

DNN deep feed-forward neural network. i, ix, xii, xiii, 45, 52, 61, 81, 83–85, 87, 93–96, 101–104, 110, 112, 114–116, 119, 121, 123–128, 130

eGeMAPS extended Geneva minimalistic acoustic parameter set. xiii, 37, 38, 81, 87, 93, 94, 99, 111–116, 128

F0 fundamental frequency. 37

FS Feature Selection. xii, 87, 102

GeMAPS Geneva minimalistic acoustic parameter set. 37, 38

GLM generalised linear model. 96, 97

GMR Gaussian mixture regression. 45, 52

GPR Gaussian process regression. 45, 46, 52

HMM hidden Markov model. 18, 38, 40

IQR inter quartile range. 80–82, 119, 121

LLD low-level descriptor. viii, xi, xiii, 4, 20, 37–39, 47, 63, 65–77, 79–81, 84, 93, 94, 99, 111, 113, 114, 116

LOSO-CV leave-one-subject-out cross-validation. 20–22, 38

LR linear regression. 46, 47, 52, 80–82, 97, 130

LSTM-RNN long short-term memory recurrent neural network. viii, 24, 28, 31–34, 37, 41, 42, 44, 45, 47, 50, 52, 59, 61, 115, 116, 124, 128

MFCC Mel-frequency cepstral coefficients. 37–39

MI mutual information. viii, 43, 67, 69–71, 73–75, 77, 82, 83, 87, 95, 100, 102, 108, 117–121, 125, 130

ML machine learning. 2, 50, 54, 63, 116

mRMR minimum redundancy maximum relevance. 43, 51, 82, 83, 87, 102, 125, 130

MSE mean squared error. vii, 25, 35

MTL multi-task learning. xii, 44, 45, 48, 52, 99, 110–112

OA output-associative. 47, 93

OCC Ortony, Clore and Collins. vii, 12, 14, 16, 17

openSMILE open speech and music interpretation by large-space extraction. 81, 93, 130

PCA Principal component analysis. 43, 51

PLLR phone log-likliehood ratio. 38

PLS Partial least squares. 46, 52

RECOLA REmote COLlaborative and Affective interactions. viii–xiii, 39, 41, 42, 46, 48–50, 54–60, 64–74, 76, 77, 79, 82–90, 93, 95, 100–114, 116–126, 128, 129

RNN recurrent neural network. vii, 28–31, 33, 34, 50, 98, 116

RVM Relevance vector machine. 45–47, 52

SAL sensitive artificial listener. 56, 57

SD standard deviation. xii, 4, 37, 40, 41, 45, 80–82, 94, 103, 105, 106, 108, 110, 117, 121, 130

SEMAINE Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression. viii–xiii, 50, 54, 56–60, 64, 66, 67, 69–74, 76, 77, 82–90, 93, 95, 100–126, 128, 129

SFS sequential forward selection. 43, 44

SSRM single-speaker regression model. 46

SVM support vector machine. 20–22, 25

SVR support vector regression. vii, 18, 24–27, 39–42, 44–47, 50, 52, 59, 61, 115, 116

TFL-MSR teacher-forced learning with multi-stage regression. i, xii, xiii, 7, 9, 93, 99, 100, 110–114, 123, 124, 126–128

ZCR zero-crossing rate. 37, 80, 81, 119, 130

Bibliography

[1] R. Descartes, *The Passions of the Soul and Other Late Philosophical Writings*, English, trans. by M. Moriarty. Oxford, United Kingdom: OUP Oxford, Nov. 2015, ISBN: 978-0-19-968413-7.

[2] C. Darwin, *The expression of the emotions in man and animals*. London, England: John Murray, 1872. DOI: 10.1037/10001-000.

[3] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie and M. Pantic, 'AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge', in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16, New York, NY, USA: ACM, 2016, pp. 3–10, ISBN: 978-1-4503-4516-3. DOI: 10.1145/2988257.2988258. [Online]. Available: http://doi.acm.org/10.1145/2988257.2988258.

[4] G. Stratou and L. P. Morency, 'MultiSense: Context-Aware Nonverbal Behavior Analysis Framework: A Psychological Distress Use Case', *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 190–203, Apr. 2017, ISSN: 1949-3045. DOI: 10.1109/TAFFC.2016.2614300.

[5] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker and M. Breakspear, 'Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors', *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 478–490, Oct. 2018, ISSN: 1949-3045. DOI: 10.1109/TAFFC.2016.2634527.

[6] J. O. Egede, S. Song, T. A. Olugbade, C. Wang, A. Williams, H. Meng, M. Aung, N. D. Lane, M. Valstar and N. Bianchi-Berthouze, 'EMOPAIN Challenge 2020: Multimodal Pain Evaluation from Facial and Bodily Expressions', Jan. 2020. arXiv: 2001.07739. [Online]. Available: http://arxiv.org/abs/2001.07739.

[7] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie and M. Pantic, 'AVEC 2014: 3d Dimensional Affect and Depression Recognition Challenge', in *Proceedings of the 4th International Workshop on*

Audio/Visual Emotion Challenge, ser. AVEC '14, New York, NY, USA: ACM, 2014, pp. 3–10, ISBN: 978-1-4503-3119-7. DOI: 10.1145/2661806.2661807. [Online]. Available: http://doi.acm.org/10.1145/2661806.2661807.

[8] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie and M. Pantic, 'AV+EC 2015: The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data', in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '15, New York, NY, USA: ACM, 2015, pp. 3–8, ISBN: 978-1-4503-3743-4. DOI: 10.1145/2808196.2811642. [Online]. Available: http://doi.acm.org/10.1145/2808196.2811642.

[9] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, E. Çiftçi, H. Güleç, A. A. Salah and M. Pantic, 'AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition', in *Proceedings of the 8th International Workshop on Audio/Visual Emotion Challenge, AVEC'18, co-located with the 26th ACM International Conference on Multimedia, MM 2018*, F. Ringeval, B. Schuller, M. Valstar, R. Cowie and M. Pantic, Eds., Seoul, Korea: ACM, Oct. 2018.

[10] L. F. Barrett, 'Valence is a basic building block of emotional life', *Journal of Research in Personality*, Proceedings of the 2005 Meeting of the Association of Research in Personality, vol. 40, no. 1, pp. 35–55, Feb. 2006, ISSN: 0092-6566. DOI: 10.1016/j.jrp.2005.08.006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0092656605000590.

[11] G. McKeown, M. Valstar, R. Cowie, M. Pantic and M. Schroder, 'The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent', *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, Jan. 2012, ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2011.20.

[12] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne and B. Schuller, 'Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data', *Pattern Recognition Letters*, Pattern Recognition in Human Computer Interaction, vol. 66, pp. 22–30, Nov. 2015, ISSN: 0167-8655. DOI: 10.1016/j.patrec.2014.11.007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865514003572.

[13] L. He, D. Jiang, L. Yang, E. Pei, P. Wu and H. Sahli, 'Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory

Recurrent Neural Networks', in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '15, New York, NY, USA: ACM, 2015, pp. 73–80, ISBN: 978-1-4503-3743-4. DOI: 10.1145/2808196.2811641. [Online]. Available: http://doi.acm.org/10.1145/2808196.2811641.

[14] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli and T. S. Huang, 'Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction', in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16, New York, NY, USA: ACM, 2016, pp. 97–104, ISBN: 978-1-4503-4516-3. DOI: 10.1145/2988257.2988264. [Online]. Available: http://doi.acm.org/10.1145/2988257.2988264.

[15] J. Han, Z. Zhang, N. Cummins, F. Ringeval and B. Schuller, 'Strength modelling for real-worldautomatic continuous affect recognition from audiovisual signals', en, *Image and Vision Computing*, Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing, vol. 65, pp. 76–86, Sep. 2017, ISSN: 0262-8856. DOI: 10.1016/j.imavis.2016.11.020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885616302177.

[16] F. Ringeval, A. Sonderegger, J. Sauer and D. Lalanne, 'Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions', in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Apr. 2013, pp. 1–8. DOI: 10.1109/FG.2013.6553805.

[17] J. A. Russell and L. F. Barrett, 'Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant', English, *Journal of Personality and Social Psychology*, vol. 76, no. 5, pp. 805–819, May 1999, ISSN: 0022-3514. DOI: http://dx.doi.org.ezproxy.ait.ie/10.1037/0022-3514.76.5.805. [Online]. Available: http://search.proquest.com/psycarticles/docview/614336643/abstract/688C997352B949CAPQ/1.

[18] R. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, en. Oxford University Press, 1997, ISBN: 978-0-19-510446-2.

[19] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller and S. S. Narayanan, 'The interspeech 2010 paralinguistic challenge', in *Eleventh Annual Conference of the International Speech Communication Association*, 2010, pp. 2794–2797.

[20] A. Cullen, J. Kane, T. Drugman and N. Harte, 'Creaky voice and the classification of affect', 2013.

[21] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan and K. P. Truong, 'The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing', *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, Apr. 2016, ISSN: 1949-3045. DOI: 10.1109/TAFFC.2015.2457417.

[22] R. B. Adams Jr. and R. E. Kleck, 'Perceived gaze direction and the processing of facial displays of emotion', *Psychological Science (0956-7976)*, vol. 14, no. 6, pp. 644–647, Nov. 2003, ISSN: 09567976. [Online]. Available: http://ezproxy.ait.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true%5C&AuthType=ip,shib%5C&db=s3h%5C&AN=11301995%5C&site=eds-live.

[23] R. B. Adams and R. E. Kleck, 'Effects of Direct and Averted Gaze on the Perception of Facially Communicated Emotion', English, *Emotion*, vol. 5, no. 1, pp. 3–11, Mar. 2005, ISSN: 1528-3542. DOI: http://dx.doi.org.ezproxy.ait.ie/10.1037/1528-3542.5.1.3. [Online]. Available: http://search.proquest.com/psycinfo/docview/614426648/abstract/EA898B0E0208425EPQ/1.

[24] A. Adams, M. Mahmoud, T. Baltrušaitis and P. Robinson, 'Decoupling facial expressions and head motions in complex emotions', in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, Sep. 2015, pp. 274–280. DOI: 10.1109/ACII.2015.7344583.

[25] C. Busso, Z. Deng, M. Grimm, U. Neumann and S. Narayanan, 'Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis', *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1075–1086, Mar. 2007, ISSN: 1558-7916. DOI: 10.1109/TASL.2006.885910.

[26] S. R. Livingstone and C. Palmer, 'Head movements encode emotions during speech and song', English, *Emotion*, vol. 16, no. 3, pp. 365–380, Apr. 2016, ISSN: 1528-3542. DOI: http://dx.doi.org.ezproxy.ait.ie/10.1037/emo0000106. [Online]. Available: http://search.proquest.com/psycarticles/docview/1727655529/abstract/FBF0D271196D40BEPQ/1.

[27] E. H. Hess and J. M. Polt, 'Pupil size as related to interest value of visual stimuli', English, *Science*, vol. 132, pp. 349–350, 1960, ISSN: 0036-8075. DOI: http://dx.doi.org.ezproxy.ait.ie/10.1126/science.132.3423.349.

[Online]. Available: http://search.proquest.com/psycinfo/docview/615377018/26A28B64DF04655PQ/34.

[28] J. M. Polt and E. H. Hess, 'Changes in pupil size to visually presented words', English, *Psychonomic Science*, vol. 12, no. 8, pp. 389–390, 1968, ISSN: 0033-3131. DOI: http://dx.doi.org.ezproxy.ait.ie/10.3758/BF03331368. [Online]. Available: http://search.proquest.com/psycinfo/docview/615571693/26A28B64DF04655PQ/35.

[29] D. Kahneman, W. S. Peavler and L. Onuska, 'Effects of verbalization and incentive on the pupil response to mental activity', English, *Canadian Journal of Psychology/Revue canadienne de psychologie*, vol. 22, no. 3, pp. 186–196, 1968, ISSN: 0008-4255. DOI: http://dx.doi.org.ezproxy.ait.ie/10.1037/h0082759. [Online]. Available: http://search.proquest.com/psycarticles/docview/614261990/abstract/5254DB3C86247F1PQ/1.

[30] L. F. Barrett, 'Seeing Fear: It's All in the Eyes?', *Trends in Neurosciences*, vol. 41, no. 9, pp. 559–563, Sep. 2018, ISSN: 0166-2236. DOI: 10.1016/j.tins.2018.06.009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016622361830170X.

[31] B. W. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. R. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente and S. Kim, 'The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism', in *Interspeech*, 2013, pp. 148–152.

[32] M. Soleymani, J. Lichtenauer, T. Pun and M. Pantic, 'A Multimodal Database for Affect Recognition and Implicit Tagging', *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, Jan. 2012, ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2011.25.

[33] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller and S. Zafeiriou, 'Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network', in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 5200–5204. DOI: 10.1109/ICASSP.2016.7472669.

[34] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, 'End-to-End Multimodal Emotion Recognition Using Deep Neural Networks', *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017, ISSN: 1932-4553. DOI: 10.1109/JSTSP.2017.2764438. [Online]. Available: http://ieeexplore.ieee.org/document/8070966/.

[35] H. Gunes, B. Schuller, M. Pantic and R. Cowie, 'Emotion representation, analysis and synthesis in continuous space: A survey', in *Face and Gesture 2011*, Mar. 2011, pp. 827–834. DOI: 10.1109/FG.2011.5771357.

[36] H. Kaya, D. Fedotov, A. Yeşilkanat, O. Verkholyak, Y. Zhang and A. Karpov, 'LSTM Based Cross-corpus and Cross-task Acoustic Emotion Recognition', en, in *Interspeech 2018*, ISCA, Sep. 2018, pp. 521–525. DOI: 10.21437/Interspeech.2018-2298. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2018/abstracts/2298.html.

[37] J. Panksepp, 'Affective neuroscience of the emotional BrainMind: Evolutionary perspectives and implications for understanding depression', *Dialogues Clin Neurosci*, vol. 12, no. 4, pp. 533–545, Dec. 2010, ISSN: 1294-8322. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3181986/.

[38] P. Ekman and W. V. Friesen, 'Constants across cultures in the face and emotion', English, *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971, ISSN: 0022-3514. DOI: http://dx.doi.org.ezproxy.ait.ie/10.1037/h0030377. [Online]. Available: http://search.proquest.com/psycarticles/docview/614303621/abstract/4A5663E3EB254A2BPQ/1.

[39] P. Ekman, 'Are there basic emotions?', English, *Psychological Review*, vol. 99, no. 3, pp. 550–553, Jul. 1992, ISSN: 0033-295X. DOI: http://dx.doi.org.ezproxy.ait.ie/10.1037/0033-295X.99.3.550. [Online]. Available: http://search.proquest.com/psycarticles/docview/614332316/abstract/DC3FDC2B303E4211PQ/4.

[40] A. Ortony, G. L. Clore and A. Collins, *The Cognitive Structure of Emotions*. Cambridge University Press, Jul. 1988. DOI: 10.1017/cbo9780511571299.

[41] K. R. Scherer, A. Schorr and T. Johnstone, *Appraisal processes in emotion : theory, methods, research*. Oxford University Press, 2001, p. 478, ISBN: 9780195130072.

[42] J. A. Russell, 'A circumplex model of affect', English, *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980, ISSN: 0022-3514. DOI: http://dx.doi.org.ezproxy.ait.ie/10.1037/h0077714. [Online]. Available: http://search.proquest.com/psycarticles/docview/614369376/abstract/DBF6BE4A22B54253PQ/1.

[43] L. F. Barrett, '2 - Navigating the Science of Emotion', en, in *Emotion Measurement*, H. L. Meiselman, Ed., Woodhead Publishing, Jan. 2016, pp. 31–63, ISBN: 978-0-08-100508-8. DOI: 10.1016/B978-0-08-100508-8.00002-3.

[Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780081005088000023.

[44] S. S. Tomkins and R. McCarter, 'WHAT AND WHERE ARE THE PRIMARY AFFECTS?SOME EVIDENCE FOR A THEORY.', *Perceptual and motor skills*, vol. 18, pp. 119–58, Feb. 1964, ISSN: 0031-5125. DOI: 10.2466/pms.1964.18.1.119. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14116322.

[45] D. J. Anderson and R. Adolphs, *A framework for studying emotions across species*, Mar. 2014. DOI: 10.1016/j.cell.2014.03.003.

[46] P. Ekman, 'Facial expression and emotion', *American Psychologist*, vol. 48, no. 4, pp. 384–392, 1993, ISSN: 0003066X. DOI: 10.1037/0003-066X.48.4.384.

[47] A. Ortony and T. J. Turner, 'What's basic about basic emotions?', *Psychological Review*, vol. 97, no. 3, pp. 315–331, 1990, ISSN: 0033295X. DOI: 10.1037/0033-295X.97.3.315.

[48] D. Sander, D. Grandjean and K. R. Scherer, 'A systems approach to appraisal mechanisms in emotion', *Neural Networks*, vol. 18, no. 4, pp. 317–352, 2005, ISSN: 08936080. DOI: 10.1016/j.neunet.2005.03.001.

[49] G. L. Clore and A. Ortony, 'Psychological construction in the OCC model of emotion', *Emotion Review*, vol. 5, no. 4, pp. 335–343, 2013, ISSN: 17540739. DOI: 10.1177/1754073913489751.

[50] W. Wundt, *Outlines of Psychology*, C. Judd, Ed., ser. Outlines of Psychology. Oxford, England: Engelmann, 1907.

[51] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch and P. C. Ellsworth, 'The world of emotions is not two-dimensional', eng, *Psychol Sci*, vol. 18, no. 12, pp. 1050–1057, Dec. 2007, ISSN: 0956-7976. DOI: 10.1111/j.1467-9280.2007.02024.x.

[52] A. S. Cowen and D. Keltner, 'Self-report captures 27 distinct categories of emotion bridged by continuous gradients', en, *PNAS*, vol. 114, no. 38, E7900–E7909, Sep. 2017, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1702247114. [Online]. Available: http://www.pnas.org/content/114/38/E7900.

[53] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. W. Schuller, K. Star, E. Hajiyev and M. Pantic, 'SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild', *IEEE Transactions on Pattern Analysis and*

Machine Intelligence, pp. 1–1, Oct. 2019, ISSN: 0162-8828. DOI: 10.1109/tpami.2019.2944808. arXiv: 1901.02839.

[54] H. Schlosberg, 'The description of facial expressions in terms of two dimensions', English, *Journal of Experimental Psychology*, vol. 44, no. 4, pp. 229–237, Oct. 1952, ISSN: 0022-1015. DOI: http://dx.doi.org.ezproxy.ait.ie/10.1037/h0055778. [Online]. Available: http://search.proquest.com/psycarticles/docview/614336281/abstract/AD37642E00294CE2PQ/1.

[55] A. B. Satpute, P. A. Kragel, L. F. Barrett, T. D. Wager and M. Bianciardi, 'Deconstructing arousal into wakeful, autonomic and affective varieties', *Neuroscience Letters*, Functional {Neuroimaging} of the {Emotional} {Brain}, vol. 693, pp. 19–28, Feb. 2019, ISSN: 0304-3940. DOI: 10.1016/j.neulet.2018.01.042. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S030439401830048X.

[56] J. Posner, J. A. Russell and B. S. Peterson, 'The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology', *Dev Psychopathol*, vol. 17, no. 3, pp. 715–734, 2005, ISSN: 0954-5794. DOI: 10.1017/S0954579405050340. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2367156/.

[57] J. Gratch and S. C. Marsella, 'Appraisal Models', in *The Oxford Handbook of Affective Computing*, R. Calvo, S. K. D'Mello, J. Gratch and A. Kappas, Eds., Oxford University Press, Jul. 2014. DOI: 10.1093/oxfordhb/9780199942237.013.015.

[58] R. Adolphs, L. Mlodinow and L. F. Barrett, *What is an emotion?*, Oct. 2019. DOI: 10.1016/j.cub.2019.09.008.

[59] J. F. Cohn, L. I. Reed, T. Moriyama, J. Xiao, K. Schmidt and Z. Ambadar, 'Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles', in *Proceedings - Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 129–135, ISBN: 0769521223. DOI: 10.1109/AFGR.2004.1301520.

[60] T. Shibata, A. Michishita and N. Bianchi-Berthouze, 'Analysis and Modelling of Affective Japanese Sitting Postures by Japanese and British Observers', in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, ISSN: 2156-8103, Sep. 2013, pp. 91–96. DOI: 10.1109/ACII.2013.22.

[61] A. Franco, C. M. Neves, C. Quintão, R. Vigário and P. Vieira, 'Singular Spectrum Analysis of Pupillometry Data. Identification of the Sympathetic and Parasympathetic Activity', *Procedia Technology*, Conference on Electronics, Telecommunications and Computers – CETC 2013. Vol. 17, pp. 273–280, Jan. 2014, ISSN: 2212-0173. DOI: 10.1016/j.protcy.2014.10.237. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2212017314004733.

[62] Z. Hammal and J. F. Cohn, 'Intra- and Interpersonal Functions of Head Motion in Emotion Communication', in *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*, Instanbul, Turkey: Association for Computing Machinery (ACM), 2014, pp. 19–22. DOI: 10.1145/2666253.2666258.

[63] P. Ricciardelli, L. Lugli, A. Pellicano, C. Iani and R. Nicoletti, 'Interactive effects between gaze direction and facial expression on attentional resources deployment: The task instruction and context matter', en, *Scientific Reports*, vol. 6, p. 21 706, Feb. 2016, ISSN: 2045-2322. DOI: 10.1038/srep21706. [Online]. Available: https://www.nature.com/articles/srep21706.

[64] M. Schneider, L. Leuchs, M. Czisch, P. G. Sämann and V. I. Spoormaker, 'Disentangling reward anticipation with simultaneous pupillometry / fMRI', *NeuroImage*, vol. 178, pp. 11–22, Sep. 2018, ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2018.04.078. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811918303987.

[65] Y. Ding, L. Shi and Z. Deng, 'Low-level Characterization of Expressive Head Motion through Frequency Domain Analysis', *IEEE Transactions on Affective Computing*, pp. 1–1, 2018, ISSN: 1949-3045. DOI: 10.1109/TAFFC.2018.2805892.

[66] N. Sadoughi, Y. Liu and C. Busso, 'Meaningful head movements driven by emotional synthetic speech', *Speech Communication*, vol. 95, pp. 87–99, Dec. 2017, ISSN: 0167-6393. DOI: 10.1016/j.specom.2017.07.004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639317300055.

[67] O. Lappi, 'Eye movements in the wild: Oculomotor control, gaze behavior & frames of reference', *Neuroscience & Biobehavioral Reviews*, vol. 69, pp. 49–68, Oct. 2016, ISSN: 0149-7634. DOI: 10.1016/j.neubiorev.2016.06.006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0149763415301317.

[68] R. J. Itier and M. Batty, 'Neural bases of eye and gaze processing: The core of social cognition', *Neuroscience & Biobehavioral Reviews*, vol. 33, no. 6, pp. 843–863, Jun. 2009, ISSN: 0149-7634. DOI: 10.1016/j.neubiorev.2009.02.004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0149763409000207.

[69] A. R. Bentivoglio, S. B. Bressman, E. Cassetta, D. Carretta, P. Tonali and A. Albanese, 'Analysis of blink rate patterns in normal subjects', eng, *Mov. Disord.*, vol. 12, no. 6, pp. 1028–1034, Nov. 1997, ISSN: 0885-3185. DOI: 10.1002/mds.870120629.

[70] R. E. Kaliouby and P. Robinson, 'Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures', in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, Jun. 2004, pp. 154–154. DOI: 10.1109/CVPR.2004.153.

[71] S. Duncan and L. F. Barrett, 'The role of the amygdala in visual awareness', *Trends in Cognitive Sciences*, vol. 11, no. 5, pp. 190–192, May 2007, ISSN: 1364-6613. DOI: 10.1016/j.tics.2007.01.007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364661307000654.

[72] H. Gunes and M. Pantic, 'Dimensional Emotion Prediction from Spontaneous Head Gestures for Interaction with Sensitive Artificial Listeners', en, in *Intelligent Virtual Agents*, J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud and A. Safonova, Eds., vol. 6356, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 371–377, ISBN: 978-3-642-15892-6. DOI: 10.1007/978-3-642-15892-6_39. [Online]. Available: http://link.springer.com/10.1007/978-3-642-15892-6_39.

[73] K. Friston, R. A. Adams, L. Perrinet and M. Breakspear, 'Perceptions as Hypotheses: Saccades as Experiments', *Front Psychol*, vol. 3, May 2012, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2012.00151. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3361132/.

[74] C. A. Cushing, H. Y. Im, R. B. Adams, N. Ward, D. N. Albohn, T. G. Steiner and K. Kveraga, 'Neurodynamics and connectivity during facial fear perception: The role of threat exposure and signal congruity', En, *Scientific Reports*, vol. 8, no. 1, p. 2776, Feb. 2018, ISSN: 2045-2322. DOI: 10.1038/s41598-018-20509-8. [Online]. Available: https://www.nature.com/articles/s41598-018-20509-8.

[75] Y. Zhao, X. Wang and E. M. Petriu, 'Facial expression anlysis using eye gaze information', in *2011 IEEE International Conference on Computational In-*

telligence for Measurement Systems and Applications (CIMSA) Proceedings, Sep. 2011, pp. 1–4. DOI: 10.1109/CIMSA.2011.6059936.

[76] F. Ringeval, A. Sonderegger, B. Noris, A. Billard, J. Sauer and D. Lalanne, 'On the Influence of Emotional Feedback on Emotion Awareness and Gaze Behavior', in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Sep. 2013, pp. 448–453. DOI: 10.1109/ACII.2013.80.

[77] J. Wang, M. X. Huang, G. Ngai and H. V. Leong, 'Are you stressed? Your eyes and the mouse can tell', in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, Oct. 2017, pp. 222–228. DOI: 10.1109/ACII.2017.8273604.

[78] S. D. Kreibig, 'Autonomic nervous system activity in emotion: A review', *Biological Psychology*, The biopsychology of emotion: Current theoretical and empirical perspectives, vol. 84, no. 3, pp. 394–421, Jul. 2010, ISSN: 0301-0511. DOI: 10.1016/j.biopsycho.2010.03.010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301051110000827.

[79] R. H. Spector, 'The Pupils', in *Clinical Methods: The History, Physical, and Laboratory Examinations*, H. K. Walker, W. D. Hall and J. W. Hurst, Eds., 3rd, Boston: Butterworths, 1990, ISBN: 978-0-409-90077-4. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK381/.

[80] E. H. Hess and J. M. Polt, 'Pupil Size in Relation to Mental Activity during Simple Problem-Solving', *Science*, vol. 143, no. 3611, pp. 1190–1192, 1964, ISSN: 0036-8075. [Online]. Available: https://www.jstor.org/stable/1712692.

[81] D. Kahneman and J. Beatty, 'Pupil Diameter and Load on Memory', *Science*, vol. 154, no. 3756, pp. 1583–1585, 1966, ISSN: 0036-8075. [Online]. Available: https://www.jstor.org/stable/1720478.

[82] C. Aracena, S. Basterrech, V. Snáel and J. Velásquez, 'Neural Networks for Emotion Recognition Based on Eye Tracking Data', in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 2015, pp. 2632–2637. DOI: 10.1109/SMC.2015.460.

[83] E. Wood, T. Baltruaitis, X. Zhang, Y. Sugano, P. Robinson and A. Bulling, 'Rendering of Eyes for Eye-Shape Registration and Gaze Estimation', in *2015 IEEE International Conference on Computer Vision (ICCV)*, ISSN: 2380-7504, Dec. 2015, pp. 3756–3764. DOI: 10.1109/ICCV.2015.428.

[84] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J. N. Antons, K. Y. Chan, N. Ramzan and K. Brunnstrom, 'Psychophysiology-Based QoE Assessment: A Survey', *IEEE Journal of Selected Topics in Signal Processing*, vol. PP, no. 99, pp. 1–1, 2016, ISSN: 1932-4553. DOI: 10.1109/JSTSP.2016.2609843.

[85] A. Kapoor, W. Burleson and R. W. Picard, 'Automatic prediction of frustration', en, *International Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 724–736, Aug. 2007, ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2007.02.003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1071581907000377.

[86] G. A. Ramirez, T. Baltrušaitis and L.-P. Morency, 'Modeling Latent Discriminative Dynamic of Multi-dimensional Affective Signals', en, in *Affective Computing and Intelligent Interaction*, S. D'Mello, A. Graesser, B. Schuller and J.-C. Martin, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2011, pp. 396–406, ISBN: 978-3-642-24571-8. DOI: 10.1007/978-3-642-24571-8_51.

[87] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie and M. Pantic, 'AVEC 2011–The First International Audio/Visual Emotion Challenge', en, in *Affective Computing and Intelligent Interaction*, S. D'Mello, A. Graesser, B. Schuller and J.-C. Martin, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2011, pp. 415–424, ISBN: 978-3-642-24571-8. DOI: 10.1007/978-3-642-24571-8_53.

[88] S. Wu, Z. Du, W. Li, D. Huang and Y. Wang, 'Continuous Emotion Recognition in Videos by Fusing Facial Expression, Head Pose and Eye Gaze', in *2019 International Conference on Multimodal Interaction*, ser. ICMI '19, Suzhou, China: Association for Computing Machinery, Oct. 2019, pp. 40–48, ISBN: 978-1-4503-6860-5. DOI: 10.1145/3340555.3353739. [Online]. Available: https://doi.org/10.1145/3340555.3353739.

[89] S. Park, S. Scherer, J. Gratch, P. Carnevale and L. P. Morency, 'Mutual Behaviors during Dyadic Negotiation: Automatic Prediction of Respondent Reactions', in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Sep. 2013, pp. 423–428. DOI: 10.1109/ACII.2013.76.

[90] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt and M. Pantic, 'AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge', in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ser. AVEC

'17, New York, NY, USA: ACM, 2017, pp. 3–9, ISBN: 978-1-4503-5502-5. DOI: 10.1145/3133944.3133953. [Online]. Available: http://doi.acm.org/10.1145/3133944.3133953.

[91] T. Baltrušaitis, P. Robinson and L. P. Morency, 'OpenFace: An open source facial behavior analysis toolkit', in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2016, pp. 1–10. DOI: 10.1109/WACV.2016.7477553.

[92] T. Baltrusaitis, A. Zadeh, Y. C. Lim and L. P. Morency, 'OpenFace 2.0: Facial Behavior Analysis Toolkit', in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, May 2018, pp. 59–66. DOI: 10.1109/FG.2018.00019.

[93] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Berlin, Heidelberg: Springer-Verlag, 1995, ISBN: 978-0-387-94559-0.

[94] B. E. Boser, I. M. Guyon and V. N. Vapnik, 'A Training Algorithm for Optimal Margin Classifiers', in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ser. COLT '92, New York, NY, USA: ACM, 1992, pp. 144–152, ISBN: 978-0-89791-497-0. DOI: 10.1145/130385.130401. [Online]. Available: http://doi.acm.org/10.1145/130385.130401.

[95] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012, ISBN: 0262018020.

[96] A. J. Smola and B. Schölkopf, 'A tutorial on support vector regression', *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004, ISSN: 09603174. DOI: 10.1023/B:STCO.0000035301.49549.88.

[97] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang and C.-J. Lin, 'LIBLINEAR: A Library for Large Linear Classification', *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008, ISSN: 1532-4435.

[98] M. Kächele, M. Schels and F. Schwenker, 'Inferring Depression and Affect from Application Dependent Meta Knowledge', in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '14, New York, NY, USA: ACM, 2014, pp. 41–48, ISBN: 978-1-4503-3119-7. DOI: 10.1145/2661806.2661813. [Online]. Available: http://doi.acm.org/10.1145/2661806.2661813.

[99] M. Schmitt, F. Ringeval and B. Schuller, 'At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech', in *Interspeech 2016*, Sep. 2016, pp. 495–499. DOI: 10.21437/Interspeech.2016-1124. [Online]. Available: http://www.isca-speech.org/archive/Interspeech%5C_2016/abstracts/1124.html.

[100] S. Hochreiter and J. Schmidhuber, 'Long Short-Term Memory', *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. [Online]. Available: http://www.mitpressjournals.org/doi/10.1162/neco.1997.9.8.1735.

[101] D. E. Rumelhart, G. E. Hinton and R. J. Williams, 'Learning representations by back-propagating errors', *Nature*, vol. 323, no. 6088, pp. 533–536, 1986, ISSN: 00280836. DOI: 10.1038/323533a0.

[102] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[103] A. Zhang, Z. C. Lipton, M. Li and A. J. Smola, *Dive into Deep Learning*. 2020, https://d2l.ai.

[104] Y. Bengio, P. Simard and P. Frasconi, 'Learning Long-Term Dependencies with Gradient Descent is Difficult', *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994, ISSN: 19410093. DOI: 10.1109/72.279181.

[105] L. I.-K. Lin, 'A Concordance Correlation Coefficient to Evaluate Reproducibility', *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989, ISSN: 0006-341X. DOI: 10.2307/2532051. [Online]. Available: http://www.jstor.org/stable/2532051.

[106] F. Weninger, F. Ringeval, E. Marchi and B. Schuller, 'Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio', in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16, New York, New York, USA: AAAI Press, Jul. 2016, pp. 2196–2202, ISBN: 978-1-57735-770-4.

[107] S. K. D'Mello and J. Kory, 'A Review and Meta-Analysis of Multimodal Affect Detection Systems', *ACM Computing Surveys*, vol. 47, no. 3, pp. 1–36, Apr. 2015, ISSN: 0360-0300. DOI: 10.1145/2682899. [Online]. Available: https://dl.acm.org/doi/10.1145/2682899.

[108] H. Al Osman and T. H. Falk, 'Multimodal Affect Recognition: Current Approaches and Challenges', in *Emotion and Attention Recognition Based on Biological Signals and Images*, InTech, Feb. 2017. DOI: 10.5772/65683. [Online]. Available: http://dx.doi.org/10.5772/65683.

[109] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller and S. Narayanan, 'Paralinguistics in speech and language—State-of-the-art and the challenge', *Computer Speech & Language*, Special issue on Paralinguistics in Naturalistic Speech and Language, vol. 27, no. 1, pp. 4–39, Jan. 2013, ISSN: 0885-2308. DOI: 10.1016/j.csl.2012.02.005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230812000162.

[110] A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller and C. D. Natlae, 'Continuous Estimation of Emotions in Speech by Dynamic Cooperative Speaker Models', *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2017, ISSN: 1949-3045. DOI: 10.1109/TAFFC.2016.2531664.

[111] S. B. Davis and P. Mermelstein, 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.', *IEEE Trans. Acoust Speech Signal Processing 28*, vol. 28, pp. 357–366, Sep. 1980.

[112] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro and K. R. Scherer, 'On the acoustics of emotion in audio: What speech, music, and sound have in common', *Frontiers in Psychology*, vol. 4, no. MAY, 2013, ISSN: 16641078. DOI: 10.3389/fpsyg.2013.00292.

[113] F. Eyben, F. Weninger, F. Gross and B. Schuller, 'Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor', in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13, New York, NY, USA: ACM, 2013, pp. 835–838, ISBN: 978-1-4503-2404-5. DOI: 10.1145/2502081.2502224. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502224.

[114] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani and M. Pantic, 'AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition', in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC '19, event-place: Nice, France, New York, NY, USA: ACM, 2019, pp. 3–12, ISBN: 978-1-4503-6913-8. DOI: 10.1145/3347320.3357688. [Online]. Available: http://doi.acm.org/10.1145/3347320.3357688.

[115] Z. Huang and J. Epps, 'An Investigation of Partition-based and Phonetically-aware Acoustic Features for Continuous Emotion Prediction from Speech', *IEEE Transactions on Affective Computing*, pp. 1–1, 2018, Conference Name: IEEE Transactions on Affective Computing, ISSN: 1949-3045. DOI: 10.1109/TAFFC.2018.2821135.

[116] D. Bose, T. Dang, V. Sethu, E. Ambikairajah and S. Fernando, 'A Novel Bag-of-Optimised-Clusters Front-End for Speech based Continuous Emotion Prediction', in *2019 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019*, Institute of Electrical and Electronics Engineers Inc., Sep. 2019, ISBN: 9781728138886. DOI: 10.1109/ACII.2019.8925490.

[117] J. Zhao, R. Li, J. Liang, S. Chen and Q. Jin, 'Adversarial Domain Adaption for Multi-Cultural Dimensional Emotion Recognition in Dyadic Interactions', in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC '19, event-place: Nice, France, New York, NY, USA: ACM, 2019, pp. 37–45, ISBN: 978-1-4503-6913-8. DOI: 10.1145/3347320.3357692. [Online]. Available: http://doi.acm.org/10.1145/3347320.3357692.

[118] H. Chen, Y. Deng, S. Cheng, Y. Wang, D. Jiang and H. Sahli, 'Efficient Spatial Temporal Convolutional Features for Audiovisual Continuous Affect Recognition', in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC '19, event-place: Nice, France, New York, NY, USA: ACM, 2019, pp. 19–26, ISBN: 978-1-4503-6913-8. DOI: 10.1145/3347320.3357690. [Online]. Available: http://doi.acm.org/10.1145/3347320.3357690.

[119] G. Rizos and B. Schuller, 'Modelling Sample Informativeness for Deep Affective Computing', in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2019-May, Institute of Electrical and Electronics Engineers Inc., May 2019, pp. 3482–3486, ISBN: 9781479981311. DOI: 10.1109/ICASSP.2019.8683729.

[120] S. Chen, Q. Jin, J. Zhao and S. Wang, 'Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition', in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '17, New York, NY, USA: ACM, 2017, pp. 19–26, ISBN: 978-1-4503-5502-5. DOI: 10.1145/3133944.3133949. [Online]. Available: http://doi.acm.org/10.1145/3133944.3133949.

[121] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu and M. Yang, 'Multimodal Continuous Emotion Recognition with Data Augmentation Using Recurrent Neural Networks', in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC'18, New York, NY, USA: ACM, 2018, pp. 57–64, ISBN: 978-1-4503-5983-2. DOI: 10.1145/3266302.3266304. [Online]. Available: http://doi.acm.org/10.1145/3266302.3266304.

[122] Y. Aytar, C. Vondrick and A. Torralba, 'SoundNet: Learning Sound Representations from Unlabeled Video', *Advances in Neural Information Processing Systems*, pp. 892–900, Oct. 2016. arXiv: 1610.09001. [Online]. Available: http://arxiv.org/abs/1610.09001.

[123] T. R. Almaev and M. F. Valstar, 'Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition', in *2013*

Humaine Association Conference on Affective Computing and Intelligent Interaction, ISSN: 2156-8111, Sep. 2013, pp. 356–361. DOI: 10.1109/ACII.2013.65.

[124] X. Xiong and F. D. l. Torre, 'Supervised Descent Method and Its Applications to Face Alignment', in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, ISSN: 1063-6919, Jun. 2013, pp. 532–539. DOI: 10.1109/CVPR.2013.75.

[125] F. Eyben, M. Wöllmer, M. F. Valstar, H. Gunes, B. Schuller and M. Pantic, 'String-based audiovisual fusion of behavioural events for the assessment of dimensional affect', in *Face and Gesture 2011*, Mar. 2011, pp. 322–329. DOI: 10.1109/FG.2011.5771417.

[126] M. Grimm and K. Kroschel, 'Evaluation of natural emotions using self assessment manikins', in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, Nov. 2005, pp. 381–385. DOI: 10.1109/ASRU.2005.1566530.

[127] J. Huang, Y. Li, J. Tao, Z. Lian, Z. Wen, M. Yang and J. Yi, 'Continuous Multimodal Emotion Prediction Based on Long Short Term Memory Recurrent Neural Network', in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '17, New York, NY, USA: ACM, 2017, pp. 11–18, ISBN: 978-1-4503-5502-5. DOI: 10.1145/3133944.3133946. [Online]. Available: http://doi.acm.org/10.1145/3133944.3133946.

[128] S. Khorram, M. McInnis and E. Mower Provost, 'Jointly Aligning and Predicting Continuous Emotion Annotations', *IEEE Transactions on Affective Computing*, 2019. DOI: 10.1109/taffc.2019.2917047.

[129] M. Schmitt, N. Cummins and B. W. Schuller, 'Continuous Emotion Recognition in Speech — Do We Need Recurrence?', en, in *Interspeech 2019*, ISCA, Sep. 2019, pp. 2808–2812. DOI: 10.21437/Interspeech.2019-2710. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/2710.html.

[130] A. Ouyang, T. Dang, V. Sethu and E. Ambikairajah, 'Speech Based Emotion Prediction: Can a Linear Model Work?', en, in *Interspeech 2019*, ISCA, Sep. 2019, pp. 2813–2817. DOI: 10.21437/Interspeech.2019-3149. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/3149.html.

[131] S. Amiriparian, M. Freitag, N. Cummins and B. Schuller, 'Feature selection in multimodal continuous emotion prediction', in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, Oct. 2017, pp. 30–37. DOI: 10.1109/ACIIW.2017.8272619.

[132] H. Peng, F. Long and C. Ding, 'Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, ISSN: 01628828. DOI: 10.1109/TPAMI.2005.159.

[133] S. Paul, N. Saoda, S. M. Rahman and D. Hatzinakos, 'Mutual information-based selection of audiovisual affective features to predict instantaneous emotional state', in *2016 19th International Conference on Computer and Information Technology (ICCIT)*, Dec. 2016, pp. 463–468. DOI: 10.1109/ICCITECHN.2016.7860242.

[134] H. Kaya, D. Fedotov, D. Dresvyanskiy, M. Doyran, D. Mamontov, M. Markitantov, A. A. Akdag Salah, E. Kavcar, A. Karpov and A. A. Salah, 'Predicting Depression and Emotions in the Cross-roads of Cultures, Para-linguistics, and Non-linguistics', ACM, Oct. 2019, pp. 27–35, ISBN: 978-1-4503-6913-8. DOI: 10.1145/3347320.3357691. [Online]. Available: http://dl.acm.org/citation.cfm?id=3347320.3357691.

[135] M. A. Nicolaou, H. Gunes and M. Pantic, 'Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space', *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, Apr. 2011, ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2011.9.

[136] D.-Y. Huang, W. Ding, M. Xu, H. Ming, M. Dong, X. Yu and H. Li, 'Multimodal Prediction of Affective Dimensions via Fusing Multiple Regression Techniques', en, in *Interspeech 2017*, ISCA, Aug. 2017, pp. 162–165. DOI: 10.21437/Interspeech.2017-1088. [Online]. Available: http://www.isca-speech.org/archive/Interspeech%5C_2017/abstracts/1088.html.

[137] P. Ekman, *Expression and the Nature of Emotion*, en, K. R. Scherer and P. Ekman, Eds. Hillsdale, New Jersey: Psychology Press, 1984, pp. 319–343.

[138] S. Parthasarathy and C. Busso, 'Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning', en, in *nterspeech 2017*, ISCA, Aug. 2017, pp. 1103–1107. DOI: 10.21437/Interspeech.2017-1494. [Online]. Available: http://www.isca-speech.org/archive/Interspeech%5C_2017/abstracts/1494.html.

[139] K. Sridhar, S. Parthasarathy and C. Busso, 'Role of Regularization in the Prediction of Valence from Speech', en, in *Interspeech 2018*, ISCA, Sep. 2018, pp. 941–945. DOI: 10.21437/Interspeech.2018-2508. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2018/abstracts/2508.html.

[140] M. Schmitt and B. W. Schuller, 'Deep recurrent neural networks for emotion recognition in speech', ser. Jahrestagung für Akustik (DAGA), Munich, Germany, 2018, pp. 1537–1540.

[141] T. Dang, B. Stasak, Z. Huang, S. Jayawardena, M. Atcheson, M. Hayat, P. Le, V. Sethu, R. Goecke and J. Epps, 'Investigating Word Affect Features and Fusion of Probabilistic Predictions Incorporating Uncertainty in AVEC 2017', in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '17, New York, NY, USA: ACM, 2017, pp. 27–35, ISBN: 978-1-4503-5502-5. DOI: 10.1145/3133944.3133952. [Online]. Available: http://doi.acm.org/10.1145/3133944.3133952.

[142] T. Dang, V. Sethu, J. Epps and E. Ambikairajah, 'An Investigation of Emotion Prediction Uncertainty Using Gaussian Mixture Regression', in *Interspeech 2017*, ISCA: ISCA, Aug. 2017, pp. 1248–1252. DOI: 10.21437/Interspeech.2017-512. [Online]. Available: http://www.isca-speech.org/archive/Interspeech%7B%5C_%7D2017/abstracts/0512.html.

[143] Z. Huang and J. Epps, 'An Investigation of Emotion Dynamics and Kalman Filtering for Speech-Based Emotion Prediction', en, in *Interspeech 2017*, ISCA, Aug. 2017, pp. 3301–3305. DOI: 10.21437/Interspeech.2017-1707. [Online]. Available: http://www.isca-speech.org/archive/Interspeech%5C_2017/abstracts/1707.html.

[144] J. Han, Z. Zhang, Z. Ren and B. Schuller, 'Implicit Fusion by Joint Audiovisual Training for Emotion Recognition in Mono Modality', in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2019-May, Institute of Electrical and Electronics Engineers Inc., May 2019, pp. 5861–5865, ISBN: 9781479981311. DOI: 10.1109/ICASSP.2019.8682773.

[145] S. Kaufman, S. Rosset and C. Perlich, 'Leakage in data mining: Formulation, detection, and avoidance', in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, New York, USA: ACM Press, 2011, pp. 556–563, ISBN: 9781450308137. DOI: 10.1145/2020408.2020496. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2020408.2020496.

[146] W. A. Jassim, R. Paramesran and N. Harte, 'Speech emotion classification using combined neurogram and INTERSPEECH 2010 paralinguistic challenge features', *IET Signal Processing*, vol. 11, no. 5, pp. 587–595, 2017, ISSN: 1751-9683. DOI: 10.1049/iet-spr.2016.0336.

[147] M. Grimm, K. Kroschel and S. Narayanan, 'The Vera am Mittag German audio-visual emotional speech database', in *2008 IEEE International Conference on Multimedia and Expo*, Jun. 2008, pp. 865–868. DOI: 10.1109/ICME.2008.4607572.

[148] J. O'Dwyer, R. Flynn and N. Murray, 'Continuous affect prediction using eye gaze', in *2017 28th Irish Signals and Systems Conference (ISSC)*, Jun. 2017, pp. 1–6. DOI: 10.1109/ISSC.2017.7983611.

[149] ——, 'Continuous affect prediction using eye gaze and speech', in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2017, pp. 2001–2007. DOI: 10.1109/BIBM.2017.8217968.

[150] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir and K. Karpouzis, 'The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data', en, in *Affective Computing and Intelligent Interaction*, A. C. R. Paiva, R. Prada and R. W. Picard, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2007, pp. 488–500, ISBN: 978-3-540-74889-2. DOI: 10.1007/978-3-540-74889-2_43.

[151] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey and M. Schröder, ''feeltrace': An instrument for recording perceived emotion in real time', en, in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, ISCA, 2000.

[152] F. Weninger, 'Introducing CURRENNT: The Munich Open-Source CUDA RecurREnt Neural Network Toolkit', *Journal of Machine Learning Research*, vol. 16, pp. 547–551, 2015. [Online]. Available: http://jmlr.org/papers/v16/weninger15a.html.

[153] U. Hess, R. B. Adams Jr. and R. E. Kleck, 'Looking at you or looking elsewhere: The influence of head orientation on the signal value of emotional facial expressions', *Motivation and Emotion*, vol. 31, no. 2, pp. 137–144, 2007, ISSN: 1573-6644(Electronic),0146-7239(Print). DOI: 10.1007/s11031-007-9057-x.

[154] T. Cover and J. Thomas, *Elements of Information Theory*. New York: John Wiley, 1991.

[155] D. Gabor, 'Theory of communication. Part 1: The analysis of information', *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, Nov. 1946, Conference Name: Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering. DOI: 10.1049/ji-3-2.1946.0074.

[156] I. Daubechies, 'The wavelet transform, time-frequency localization and signal analysis', *IEEE Transactions on Information Theory*, vol. 36, no. 5, pp. 961–1005, Sep. 1990, ISSN: 0018-9448. DOI: 10.1109/18.57199.

[157] S. Mallat, *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*, 3rd. USA: Academic Press, Inc., 2008, ISBN: 0123743702.

[158] I. Daubechies, 'Orthonormal bases of compactly supported wavelets', *Communications on Pure and Applied Mathematics*, vol. 41, no. 7, pp. 909–996, Oct. 1988, ISSN: 00103640. DOI: 10.1002/cpa.3160410705. [Online]. Available: http://doi.wiley.com/10.1002/cpa.3160410705.

[159] N. De Jay, S. Papillon-Cavanagh, C. Olsen, N. El-Hachem, G. Bontempi and B. Haibe-Kains, 'mRMRe: An R package for parallelized mRMR ensemble feature selection', en, *Bioinformatics*, vol. 29, no. 18, pp. 2365–2368, Sep. 2013, Publisher: Oxford Academic, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt383. [Online]. Available: https://academic.oup.com/bioinformatics/article/29/18/2365/239921.

[160] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird and B. Schuller, 'Snore Sound Classification Using Image-Based Deep Spectrum Features', in *Interspeech 2017*, vol. 2017-August, ISCA: ISCA, Aug. 2017, pp. 3512–3516. DOI: 10.21437/Interspeech.2017-434. [Online]. Available: http://www.isca-speech.org/archive/Interspeech%7B%5C_%7D2017/abstracts/0434.html.

[161] '5.3.3.4. Fractional factorial designs', in *NIST/SEMATECH e-Handbook of Statistical Methods*, National Institute of Standards and Technology, 2012. [Online]. Available: https://www.itl.nist.gov/div898/handbook/pri/section3/pri334.htm (visited on 08/06/2020).

[162] C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup and B. Bischl, 'Pitfalls to Avoid when Interpreting Machine Learning Models', in *37th International Conference on Machine Learning*, arXiv: 2007.04131, Jul. 2020. [Online]. Available: http://arxiv.org/abs/2007.04131.

[163] C. Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019, ISSN: 25225839. DOI: 10.1038/s42256-019-0048-x.

[164] T. C. Schneirla, 'An evolutionary and developmental theory of biphasic processes underlying approach and withdrawal', in *Nebraska symposium on motivation, 1959*, Oxford, England: Univer. Nebraska Press, 1959, pp. 1–42.

[165] G. Greenberg, *Approach/Withdrawal Theory*. Springer International Publishing, 2017, pp. 1–6. DOI: 10.1007/978-3-319-47829-6_1074-1.

[166] A. Jain, R. Bansal, A. Kumar and K. Singh, 'A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students', *International Journal of Applied and Basic Medical Research*, vol. 5, no. 2, p. 124, 2015, ISSN: 2229-516X. DOI: 10.4103/2229-516x.157168.

[167] J. Gu, L. Cao and B. Liu, 'Modality-general representations of valences perceived from visual and auditory modalities', *NeuroImage*, vol. 203, p. 116 199, Dec. 2019, ISSN: 10959572. DOI: 10.1016/j.neuroimage.2019.116199.

[168] E. Vatikiotis-Bateson, A. V. Barbosa, C. Y. Chow, M. Oberg, J. Tan and H. C. Yehia, 'Audiovisual lombard speech: Reconciling production and perception', *Proceedings of the International Conference on Auditory-Visual Speech Processing*, pp. 45–50, 2007.

[169] F. Eyben, S. Petridis, B. Schuller and M. Pantic, 'Audiovisual vocal outburst classification in noisy acoustic conditions', in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2012, pp. 5097–5100, ISBN: 9781467300469. DOI: 10.1109/ICASSP.2012.6289067.

[170] T. Kawahara, K. Inoue, D. Lala and K. Takanashi, 'Audio-Visual Conversation Analysis by Smart Posterboard and Humanoid Robot', in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, Institute of Electrical and Electronics Engineers Inc., Sep. 2018, pp. 6573–6577, ISBN: 9781538646588. DOI: 10.1109/ICASSP.2018.8461470.

[171] L. Zhang and R. J. Radke, 'A Multi-Stream Recurrent Neural Network for Social Role Detection in Multiparty Interactions', *IEEE Journal on Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 554–567, Mar. 2020, ISSN: 19410484. DOI: 10.1109/JSTSP.2020.2992394.

[172] R. Tibshirani and R. Tibshirani, 'Regression Shrinkage and Selection Via the Lasso', *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 58, pp. 267–288, 1994. [Online]. Available: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.7574.

[173] M. Roddy and N. Harte, 'Detecting conversational gaze aversion using unsupervised learning', in *2017 25th European Signal Processing Conference (EUSIPCO)*, ISSN: 2076-1465, Aug. 2017, pp. 76–80. DOI: 10.23919/EUSIPCO.2017.8081172.

[174] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/.